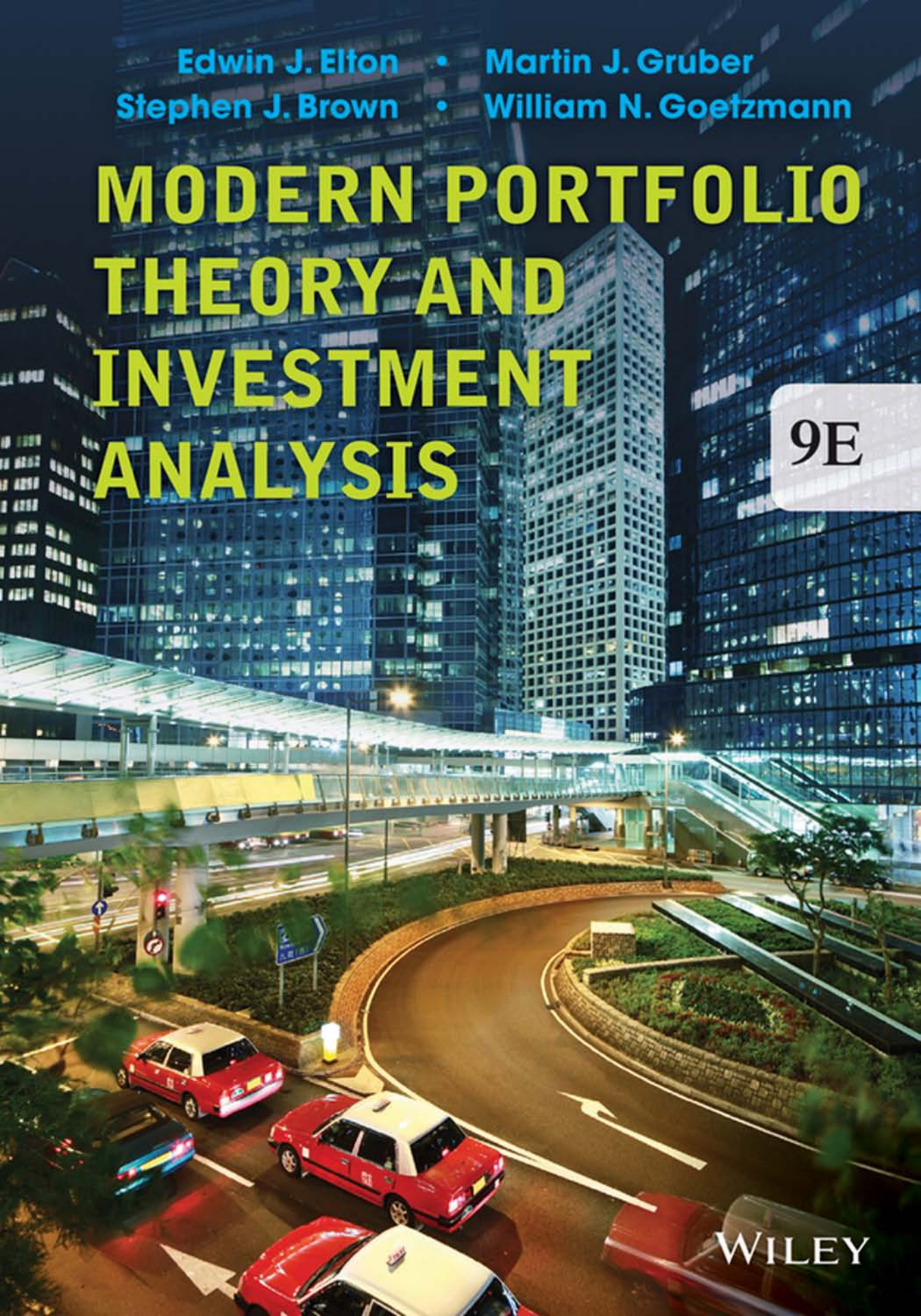
Edwin J. Elton • Martin J. Gruber
Stephen J. Brown • William N. Goetzmann

MODERN PORTFOLIO THEORY AND INVESTMENT ANALYSIS

9E



WILEY

MODERN PORTFOLIO THEORY AND INVESTMENT ANALYSIS

NINTH EDITION

EDWIN J. ELTON
Leonard N. Stern School of Business
New York University

MARTIN J. GRUBER
Leonard N. Stern School of Business
New York University

STEPHEN J. BROWN
Leonard N. Stern School of Business
New York University

WILLIAM N. GOETZMANN
School of Management Yale University

WILEY

Vice President and Executive Publisher	George Hoffman
Executive Editor	Joel Hollenbeck
Content Editor	Jennifer Manias
Assistant Editor	Courtney Luzzi
Senior Editorial Assistant	Erica Horowitz
Director of Marketing	Amy Scholz
Assistant Marketing Manager	Puja Katariwala
Marketing Assistant	Mia Brady
Senior Production Manager	Janis Soo
Associate Production Manager	Joel Balbin
Production Editor	Yee Lyn Song
Cover Designer	Kenji Ngieng
Cover Credit	© TommL/iStockphoto

This book was set in Times Roman by Thomson Digital and printed and bound by Lightning Source. The cover was printed by Lightning Source.

This book is printed on acid-free paper.

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

Library of Congress Cataloging-in-Publication Data

Elton, Edwin J.
Modern portfolio theory and investment analysis / Edwin J. Elton, Leonard N. Stern School of Business, New York University, Martin J. Gruber, Leonard N. Stern School of Business, New York University, Stephen J. Brown, Leonard N. Stern School of Business, New York University, William N. Goetzmann, Yale University.—Ninth edition.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-46994-1 (pbk.)
1. Portfolio management. 2. Investment analysis. I. Title.
HG4529.5.E47 2014
332.6–dc23

2013022155

Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

To some of the future generation of our readers: Ned's grandchildren Erik Beitel, Sophia Beitel, Miranda Beitel, Chloe Elton, Jean Paul Elton, Petra Elton, Johanna Elton, and Klara Elton, and Marty's grandchildren Samuel Gruber, Jack Gruber, and Ava Gruber.

About the Authors

Edwin J. Elton is Scholar in Residence and Professor Emeritus of Finance at the Stern School of Business of New York University. He has authored or coauthored eight books and more than 110 articles. These articles have appeared in journals such as the *Journal of Finance,* the *Review of Financial Studies, Review of Economics and Statistics, Management Science, Journal of Financial Economics, Journal of Business, Oxford Economic Papers,* and *Journal of Financial and Quantitative Analysis.* He has been coeditor of the *Journal of Finance.* Professor Elton has been a member of the board of directors of the American Finance Association and an Associate Editor of *Management Science.* Professor Elton has served as a consultant for many major financial institutions. A compendium of articles by Professor Elton and Professor Gruber has been published in two volumes by MIT Press and one volume by World Scientific Press. Professor Elton is a past president of the American Finance Association, a fellow of that association, a recipient of a distinguished research award by the Eastern Finance Association, and a recipient of the James Vertin Award from the Financial Analyst Association.

Martin J. Gruber is Scholar in Residence and Professor Emeritus of Finance, as well as past chairman of the Finance Department, at the Stern School of Business of New York University. He is a fellow of the American Finance Association. He has published nine books and more than 100 journal articles in journals such as the *Journal of Finance,* the *Review of Financial Studies, Review of Economics and Statistics, Journal of Financial Economics, Journal of Business, Management Science, Journal of Financial and Quantitative Analysis, Operations Research, Oxford Economic Papers,* and *the Journal of Portfolio Management.* He has been coeditor of the *Journal of Finance.* He has been president of the American Finance Association, a director of the European Finance Association, a director of the American Finance Association, and a director of both the Computer Applications Committee and the Investment Technology Symposium of the New York Society of Security Analysts. He was formerly Finance Department editor for *Management Science* and an Associate Editor of the *Financial Analysts Journal.* Professor Gruber has consulted in the areas of investment analysis and portfolio management with many major financial institutions. He is currently a Director of the Daiwa closed-end funds and the Aberdeen Singapore Fund. He is formerly a Director of TIAA, Director and Chairman of CREF, Director of DWS Mutual Funds, and Director of the SQ Cowen Mutual Funds.

Stephen J. Brown is David S. Loeb Professor of Finance at the Leonard N. Stern School of Business, New York University. Following successive appointments at Bell Laboratories and Yale, he joined the faculty of New York University in 1986. In 2002 he was appointed Professorial Fellow at the University of Melbourne. He has served as President of the Western Finance Association and Secretary/Treasurer of that organization, has served on

the Board of Directors of the American Finance Association, and was a founding editor of the *Review of Financial Studies*. He is a Managing Editor of *the Journal of Financial and Quantitative Analysis* and has served on the editorial board of *the Journal of Finance* and other journals. He has published numerous articles and five books on finance and economics related areas. He has served as an expert witness for the U.S. Department of Justice and testified on his research before a Full Committee Hearing of the U.S. Congress House Financial Services Committee in March 2007. In 2010 he served as a member of the Research Evaluation Committee of the Excellence in Research Australia initiative on behalf of the Commonwealth Government of Australia.

William N. Goetzmann is the Edwin J. Beinecke Professor of Finance and Management Studies and the Director of the International Center for Finance at the Yale School of Management. He has served as the president of the Western Finance Association and the European Finance Association. His published research includes work on portfolio management, investment funds, equity markets, real estate, global investing, endowment management, and the economics of the arts. He has served on the board or investment committee of various financial institutions, funds, and endowments. His other coauthored books include *The Great Mirror of Folly: Finance, Culture, and the Great Crash of 1720* (2013), *The Origins of Value: The Financial Innovations That Created Modern Capital Markets* (2005), and *The Equity Risk Premium: Essays and Explorations* (2006). He served on the Financial and Valuation Advisory Committee to the Congressional Oversight Panel to Review the Current State of Financial Markets and the Regulatory System in 2008–2009 and was the coauthor of a study on the Norwegian sovereign fund, *Evaluation of Active Management of the Norwegian Government Pension Fund—Global*, for the Norwegian Ministry of Finance in 2009.

New to the 9th Edition

There has been a renewed interest in the science of investment management in the years since the global financial crisis. The volatility of world markets and the shock to its financial institutions has caused a profound reexamination of risk, research into the methods of effective diversification, and exploration of the fundamental expected returns from financial assets. Rather than causing a rejection of modern portfolio theory, however, the financial crisis highlighted the validity of its fundamental tenants: higher expected returns require a willingness to accept higher risks; the methodology of diversification is extremely important; a longer-term perspective and an understanding of the broader scope of financial history is vital.

National and world events together with important new theoretical and empirical research have motivated a major revision of this book.

Almost all of the chapters have been revised, while more than half have been substantially rewritten. Modern developments in the theoretical and empirical literature have been incorporated into the text. All examples in the text have been brought up to date. A new chapter had been added to describe changing conditions in the mutual fund industry.

Some of the key changes in the text include the following:

- Recognizing the structural changes that have occurred in the markets in which securities are traded
- Recognizing the causes of the financial crisis of 2008 and the financial instruments that effected the crisis
- Recognizing new ways of estimating returns
- Incorporating recent developments in multiperiod consumption and investment models
- Recognizing the increased importance of international investing and diversification and the advances made in understanding emerging market investing
- Incorporating a new mode of investing: factor-based investing
- Incorporating the new theoretical and empirical literature, which helps us understand and diagnose mutual fund performance
- Incorporating new research on the efficient market theory and its origins
- Incorporating current research and applications of Bayesian methods in finance

The authors would like to thank our colleagues Joel Hasbrouck, Paul Zarowin, and Steve Figlewski for major contributions to the chapters on market structure, earnings estimation, and futures. We would also like to thank Nancy Mack and Jude Warne for assistance in preparing this manuscript.

Preface

This book, as the title suggests, is concerned with the characteristics and analysis of individual securities, as well as with the theory and practice of optimally combining securities into portfolios. Part 1 of the book provides a description of securities and markets. Two chapters provide the reader with the institutional background to place the analytics that follow in perspective.

The second, and longest, part of the book discusses modern portfolio theory. We begin Part 2 with a detailed presentation of the theory of modern portfolio analysis and show that the characteristics of portfolios are significantly different from those of the individual securities from which they are formed. In fact, portfolio analysis is the recipe for one of the few "free lunches" in economics. By the end of Chapter 6, the reader will have learned the basis of portfolio theory from the relationship of portfolio characteristics to security characteristics to the method of computing sets of portfolios that investors will find desirable.

The theory presented at the beginning of the book has been around long enough that major breakthroughs have occurred in its implementation. These breakthroughs involve simplification of the amount and type of inputs to the portfolio problem (Chapters 7 and 8), as well as simplification of the computational procedure to find sets of desirable portfolios (Chapter 9). The major advantage in the latter simplification is that the portfolio selection process and the final portfolios selected have a structure with a clear-cut economic rationale, one to which both the practicing security analyst and the economist can relate. Chapter 10 discusses the all-important input to portfolio management expected return.

The reader might note that up to now we have discussed sets of portfolios. These sets contain portfolios that would be desirable to any investor. In Chapter 11, we examine how an individual investor might choose the one optimal portfolio (for him or her) from among the sets of portfolios designed to appeal to any investor. We conclude Part 2 with a discussion of the potential benefits derived from diversifying portfolios internationally.

Part 3 provides a discussion of equilibrium in the capital markets. This material usually is included under the rubric of the capital asset pricing model or arbitrage pricing theory and shows how portfolio theory can be used to infer what equilibrium returns and prices will be for individual securities. This area is changing rapidly. But, as the reader will see, empirical tests suggest that the theory as it now stands provides great insight into the functioning of security markets and the pricing of individual issues. It also suggests ways that equilibrium theory can be used to manage portfolios more meaningfully.

Part 4 of this book deals with the characteristics and evaluation of individual securities. In this part we discuss whether security markets are efficient, the valuation of common stocks, the characteristics of earnings and their role in the valuation process, the valuation of bonds, the nature of and valuation of options, and finally the valuation and uses of futures. In addition, we explore the new field of behavioral finance and its implications for investor action and asset prices.

Part 5 is a discussion of the evaluation of the investment analysis and portfolio management process. In writing this part we have stressed techniques for evaluating every stage of the process, from the forecasting of earnings by security analysts to the performance of portfolios that are finally selected. It seems fitting that a book that deals primarily with investment analysis and portfolio management should end with a discussion of how to tell if these functions are performed well.

The book was designed to serve as a text for courses both in portfolio theory and in investment analysis that have an emphasis on portfolio theory. We have used it for these purposes at New York University for several years. For the course in portfolio analysis, we use Chapters 4–16 plus Chapters 25, 26, and 28. This thoroughly introduces the students to modern portfolio theory and general equilibrium models (capital asset pricing models and arbitrage pricing models).

The book can also be used in a course in investments where both portfolio analysis and security analysis are discussed. For these purposes, the institutional material in Chapters 1 and 2, the security analysis chapters of Part 4, as well as Chapter 26 on the evaluation of security analysis, are appropriate, and some of the advanced portfolio theory and general equilibrium chapters of Parts 2 and 3 can be deleted. Each professor's preference and the dictates of the course will ultimately determine the final choice. One possible choice that has been successfully used was the replacement of much of Chapter 6 and Chapters 8, 11, 14, 15, and 16 with the chapters on security analysis contained in Part 4. Courses covering portfolio theory and investments vary greatly in their content. We have included in this book those areas that we view as most relevant.

We believe that this book will be an aid to the practicing security analyst and portfolio manager. It is remarkable how quickly the ideas of modern portfolio theory have found their way into investment practice. The manager who wishes an overview of modern portfolio theory and investment analysis will find that Chapters 4, 5, 7, 9, 12, and 17–26 will provide a thorough and readable understanding of the issues. Specialists who are concerned with issues on implementation will find that the other chapters will equip them with the most modern tools available.

As the reader may know, New York University has not only the normal MBA and undergraduate student courses but also courses intended for full-time portfolio managers and securities analysts. The professional reader can be assured that the book has been used in these courses and that some of our most enthusiastic responses came from practicing managers who learned not only the ideas of modern portfolio theory and investment analysis but also its strengths and weaknesses.

In writing this book, our purpose has been to make all the material accessible to students of portfolio analysis and investment management, at both the undergraduate and the graduate levels. To the extent possible, the text stresses the economic intuition behind the subject matter. Mathematical proofs involving more than simple algebra are placed in footnotes, appendices, or specially noted sections of the text. They can be deleted without losing the general thrust of the subject matter. In addition, we have included problems both in the text and at the end of each chapter. We have tried to capture in this book the frontier of the state of the art of modern portfolio analysis, general equilibrium theory, and investment analysis, while presenting it in a form that is accessible and has intuitive appeal.

A book must, of necessity, present material in a certain order. We have tried to present the material so that much of it can be used in alternative sequences. For example, we tend to teach formal utility analysis after many of the concepts of portfolio analysis. However, we realize that many professors prefer to begin with a discussion of utility analysis. Thus this chapter in particular could be read immediately after the introductory chapter.

We wish to thank Professor Chris Blake for his help in preparing the problem sets included in this book.

Finally, we wish to acknowledge Dr. Watson. We have noted her contribution to utility analysis and security valuation in previous books. Her contribution to earlier versions of this book were substantial. Her untimely death meant that we did not have the benefit of her excellent advice on this latest edition, though her help is still reflected in the book you have before you.

Final Thoughts

More than 35 years have passed since we began to write the first edition of this book. Progress has been made in several areas, and yet new changes have occurred that reopen old questions. The acceptance of quantitative techniques by the investment community both here and overseas has grown at a rate we would not have dreamed of then. The use of modern portfolio techniques for stocks and bonds, dividend discount models, concepts of passive portfolios, the incorporation of international assets in portfolios, and the use of futures and options as risk control techniques are very widespread. Yet the world of investments continues to change. No sooner do we begin to believe that the capital asset pricing model (CAPM) describes reality than the arbitrage pricing theory (APT) comes along. No sooner do we convince ourselves that markets are efficient than market anomalies become hot topics. No sooner do we say that security analysis does not pay than we justify the cost of analysis in a world of partially revealing prices. No sooner is market timing discredited than it arises again under the name of tactical asset allocation.

Will the field continue to evolve and will today's truths become less true tomorrow? Probably. We will continue to learn. We know more about the capital markets now than we did 20 years ago. There is still a lot more to learn. That is why there will no doubt be a tenth edition of this book and why there are securities and strategies that have expected returns above the riskless rate.

E. J. Elton
M. J. Gruber
S. J. Brown
W. N. Goetzmann

Contents

Part 1

INTRODUCTION

1

Introduction

Almost everyone owns a portfolio (group) of assets. This portfolio is likely to contain real assets, such as a car, a house, or a refrigerator, as well as financial assets, such as stocks and bonds. The composition of the portfolio may be the result of a series of haphazard and unrelated decisions, or it may be the result of deliberate planning. In this book we discuss the basic principles underlying rational portfolio choice and what this means for prices determined in the marketplace. We confine our attention to financial assets, although much of the analysis we develop is equally applicable to real assets.

An investor is faced with a choice from among an enormous number of assets. When one considers the number of possible assets and the various possible proportions in which each can be held, the decision process seems overwhelming. In the first part of this book we analyze how decision makers can structure their problems so that they are left with a manageable number of alternatives. Later sections of the book deal with rational choice among these alternatives, methods for implementing and controlling the decision process, and equilibrium conditions in the capital markets to which the previous analysis leads.

Let us examine the composition of this book in more detail.

OUTLINE OF THE BOOK

This book is divided into five parts. The first part provides background material on securities and financial markets. The reader already familiar with these topics can go directly to Part 2.

The second and longest part deals with the subject of portfolio analysis. Portfolio analysis is concerned with finding the most desirable group of securities to hold, given the properties of each of the securities. This part of the book is itself divided into four sections. The first of these sections is titled "Mean Variance Portfolio Theory." This section deals with determining the properties of combinations (portfolios) of risky assets given the properties of the individual assets, delineating the characteristics of portfolios that make them preferable to others, and, finally, showing how the composition of the preferred portfolios can be determined.

At the end of this section readers will know almost all that they need to know about the theory of portfolio selection. This theory is more than 50 years old. In the ensuing years, a tremendous amount of work has been devoted to implementing this theory. The second

section of Part 2 is concerned with the implementation and simplification of portfolio theory. The topics covered include simplifying the quantity and type of input needed to do portfolio analysis and simplifying the computational procedure used to find the composition of the efficient portfolios.

The third section of Part 2 deals with the selection of that one portfolio that best meets the needs of an investor. We discuss not only techniques that rely on utility maximization but also other techniques suggested in the literature.

The final section of Part 2 deals with the impact of the opportunity to diversify a stock portfolio across international boundaries. As the reader might suspect, any increase in the set of possible investment opportunities should increase portfolio performance.

Part 3 deals with models of equilibrium prices and returns in the capital markets. If investors behave as portfolio theory suggests they should, then their actions can be aggregated to determine prices at which securities will sell.

The first two chapters of Part 3 deal with some alternative forms of equilibrium relationships. Different assumptions about the characteristics of capital markets and the way investors behave lead to different models of equilibrium. The third chapter in this part of the book deals with empirical tests of how well these theoretical models describe reality. The final chapter in Part 3 presents both the theoretical basis of and empirical evidence on the newest theory of relative prices: the Arbitrage Pricing Theory.

The fourth part of the book deals with some issues in investment analysis. The first question examined is the speed with which new information is incorporated into the share price. If new information is immediately and accurately incorporated into the share price, then there can be no payoff from security analysis, whereas if information is more slowly incorporated into the share price, it may pay to engage in certain types of analysis. The key to security analysis is the method used to turn forecasts of fundamental firm characteristics into forecasts of price performance. This is the subject of the second chapter in Part 4, titled "The Valuation Process." Virtually every valuation process employs forecasts of earnings as one important input. A detailed analysis of earnings is presented as an example of methods of forecasting inputs to valuation models. This is followed by a chapter that discusses noneconomic behavior and the impact of this behavior on security prices. The next two chapters in Part 4 deal with the theory of interest rates, the pricing of bonds, and the management of bond portfolios. The final two chapters in Part 4 deal with the valuation of options and financial futures. The markets for security options and for futures are among the fastest-growing markets in the country. In addition, the theory of option pricing has important implications for generating the inputs to portfolio analysis. Futures, because of their low transaction costs, are an important tool for modifying portfolio composition.

The fifth part of the book is concerned with evaluating the investment process. The first chapter in this section contains a description of the principal types of mutual funds and reviews two specific types, closed-end funds and exchange-traded funds, in some detail. The second chapter deals with the evaluation of portfolio performance with an emphasis on open-end mutual funds. In this chapter we discuss the best methods of evaluating portfolio performance and how well-managed portfolios have performed. In contrast to the voluminous literature on portfolio performance, almost nothing has been written about how to evaluate the other steps in the investment process. For example, very little has been written about how to evaluate forecasts of security analysts or how to evaluate the valuation process. The third chapter in this part of the book deals with these problems. The final chapter of the book integrates the material contained in the earlier parts.

THE ECONOMIC THEORY OF CHOICE: AN ILLUSTRATION UNDER CERTAINTY

All decision problems have certain elements in common. Any problem involves the delineation of alternatives, the selection of criteria for choosing among those alternatives, and, finally, the solution of the problem. Furthermore, individual solutions can often be aggregated to describe equilibrium conditions that prevail in the marketplace. A large part of this book will be concerned with following these steps for the selection of risky assets. But before we start this problem, let us examine a simpler one, under certainty, to illustrate the elements of the solution to any economic problem.

Consider an investor who will receive with certainty an income of $10,000 in each of two years. Assume that the only investment available is a savings account yielding 5% per year. In addition, the investor can borrow money at a 5% rate.

How much should the investor save and how much should he or she consume each year? The economic theory of choice proposes to solve this problem by splitting the analysis into two parts: first, specify those options that are available to the investor; second, specify how to choose among these options. This framework for analysis carries over to more complex problems.

The Opportunity Set

The first part of the analysis is to determine the options open to the investor. One option available is to save nothing and consume $10,000 in each period. This option is indicated by the point *B* in Figure 1.1.

Scrooge would choose another option. He would save all income in the first period and consume everything in the second. In the second period his savings account would be worth the $10,000 he saves in period 1 plus interest of 5% on the $10,000, or $10,500.

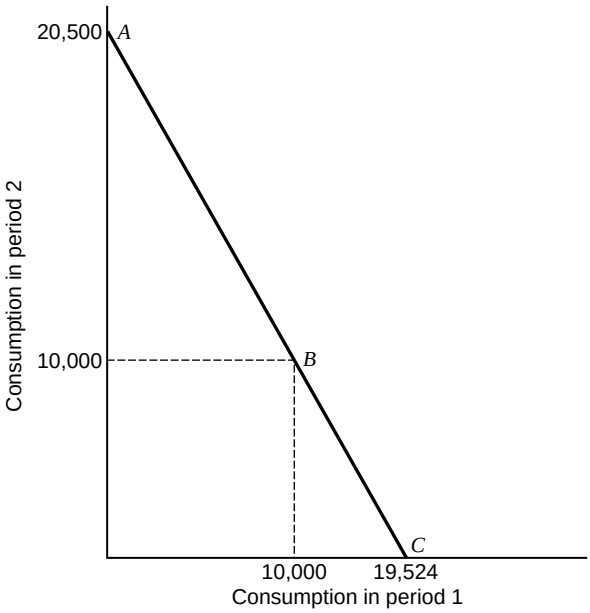


Figure 1.1 The investor's opportunity set.

Adding this to his second-period income of $10,000 gives him a consumption in period 2 of $10,500 + $10,000 = $20,500. This is indicated by point *A* in Figure 1.1.

Another possibility is to consume everything now and not worry about tomorrow. This would result in consumption of $10,000 from this period's income plus the maximum the investor could borrow against next period's income. If *X* is the amount borrowed, then *X* plus the interest paid for borrowing *X* equals the amount paid back. Because the investor's income in the second period is $10,000, the maximum amount is borrowed if *X* plus the interest on *X* at 5% equals $10,000:

$$X + 0.05X = 10{,}000$$

or

$$X = \frac{10{,}000}{1.05} = \$9{,}524$$

Thus the maximum the investor can consume in the first period is $19,524. This is indicated by point *C* in Figure 1.1. Note that points *A*, *B*, and *C* lie along a straight line. This did not happen by accident. In fact, all of the enormous possible patterns of consumption in periods 1 and 2 will lie along this straight line. Let us see why.

The amount the investor consumes in the two periods is constrained by the amount of income the investor has available in the two periods. Let C_1 be the consumption in period 1 and C_2 be the consumption in period 2. The amount consumed in period 2 is the income in period 2 of $10,000 plus the period 2 value of the savings in period 1. Remember that the value of period 1 savings can be negative, for the investor could have dissaved. That is, he could have borrowed in period 1 and consumed more than his period 1 income. As of period 2, the value of the savings in period 1 is the amount saved in period 1 ($10,000 minus what is consumed) plus accumulated interest. Putting this in equation form, we have

$$\begin{bmatrix}\text{Period 2}\\ \text{consumption}\end{bmatrix} = \begin{bmatrix}\text{Period 2}\\ \text{income}\end{bmatrix} + \begin{bmatrix}\text{Amount}\\ \text{saved in 1}\end{bmatrix}[1+0.05]$$

$$C_2 = \$10{,}000 + (10{,}000 - C_1)(1.05)$$

$$C_2 = \$20{,}500 - (1.05)C_1$$

This is, of course, the equation for a straight line and is the line shown in Figure 1.1. It has an intercept of $20,500, which results from zero consumption in period 1 ($C_1 = 0$) and is the point *A* we determined earlier. It has a slope equal to -1.05 or minus the quantity 1 plus the interest rate. The value of the slope reflects the fact that each dollar the investor consumes in period 1 is a dollar he cannot invest and, hence, reduces period 2 consumption by one dollar plus the interest he could earn on the dollar, or a total of $1.05. Thus an increase in period 1's consumption of a dollar reduces period 2's consumption by $1.05.

The investor is left with a large number of choices. We usually refer to the set of choices facing the investor as the opportunity set. Let us now examine how an investor selects the optimum consumption pattern from the opportunity set.

The Indifference Curves

The economic theory of choice states that an investor chooses among the opportunities shown in Figure 1.1 by specifying a series of curves called *utility functions* or *indifference curves*. A representative set is shown in Figure 1.2. These curves represent the investor's preference for income in the two periods. The name "indifference curves" is used because

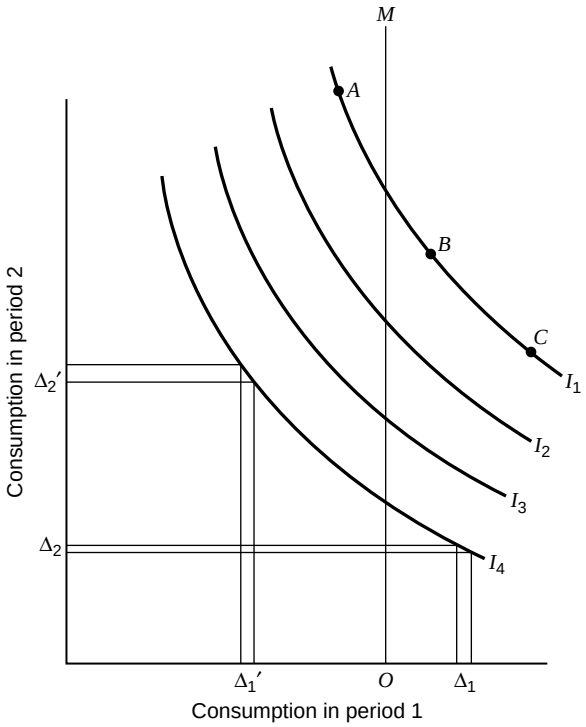


Figure 1.2 Indifference curves.

the curves are constructed so that everywhere along the same curve the investor is assumed to be equally happy. In other words, the investor does not care whether he obtains point *A*, *B*, or *C* along curve I_1.

Choices along I_1 will be preferred to choices along I_2, and choices along I_2 will be preferred to choices along I_3, and so on. This ordering results from an assumption that the investor prefers more to less. Consider the line *OM*. Along this line the amount of consumption in period 1 is held constant. As can be seen from Figure 1.2, along the line representing equal consumption in period 1, I_1 represents the most consumption in period 2, I_2 the next most, and so on. Thus, if investors prefer more to less, I_1 dominates I_2, which dominates I_3.

The curved shape results from an assumption that each additional dollar of consumption forgone in period 1 requires greater consumption in period 2. For example, if consumption in period 1 is large relative to consumption in period 2, the investor should be willing to give up a dollar of consumption in period 1 in return for a small increase in consumption in period 2. In Figure 1.2 this is illustrated by Δ_1 for the amount the investor gives up in period 1 and Δ_2 for the amount the investor gains in period 2. However, if the investor has very few dollars of consumption in period 1, then a large increase in period 2 is required to be indifferent about giving up the extra consumption in period 1. This is represented by the Δ_1' in period 1 (which is the same size as Δ_1) and the Δ_2' in period 2 (which is much larger than Δ_2).

The Solution

The indifference curves and the opportunity set represent the tools necessary for the investor to reach a solution. The optimum consumption pattern for the investor is determined by the

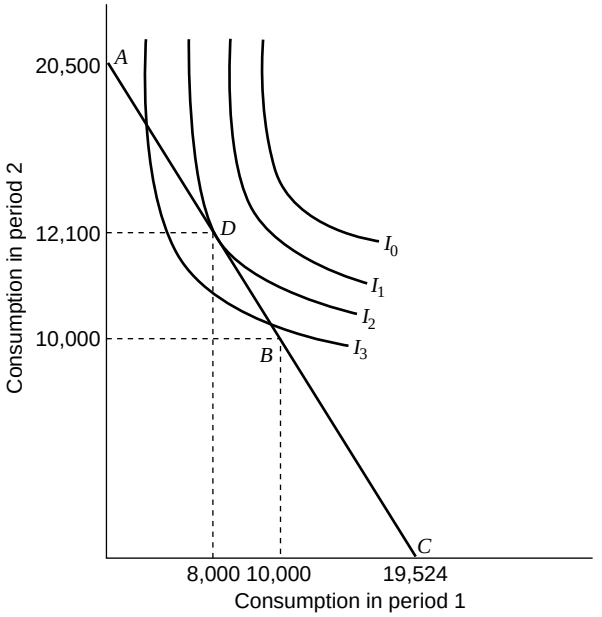


Figure 1.3 Investor equilibrium.

point at which a number of the set of indifference curves is tangent to the opportunity set (point *D* in Figure 1.3). Let us see why. The investor can select either of the two consumption patterns indicated by the points where I_3 intersects the line *ABC* in Figure 1.3. But we have argued that the investor is better off selecting a consumption pattern lying on an indifference curve located above and to the right of I_3, if possible. The investor will move to higher indifference curves until the highest one that contains a feasible consumption pattern is reached. That is the one just tangent to the opportunity set. This is I_2 in Figure 1.3, and the consumption pattern the investor will choose is given by the point of tangency, point *D*. The question might be asked, why doesn't the investor move up to a point along I_0 because this would be preferable to a point along I_2? The answer is that there is no investment opportunity available on line I_0.

An Example: Determining Equilibrium Interest Rates

We take another look at the investor's possible decision to see how it can help in determining equilibrium conditions in the market. The optimum decision could occur in three sections of Figure 1.3: *A* to *B*, point *B*, or *B* to *C*. If the optimum occurs in the segment *AB*, then the investor lends money at the 5% rate. If the optimum occurs at point *B*, then the investor is neither a borrower nor a lender. Finally, if the optimum occurs in segment *BC*, then the investor borrows against future income at the 5% rate.

In this simple framework, equilibrium in the marketplace is easy to determine. At a 5% interest rate this investor wishes to lend $2,000, the difference between $10,000 in income and $8,000 in consumption. Summing across all investors who wish to lend when the interest rate is 5% gives one point on the supply curve. Similarly, summing across investors who wish to borrow at a 5% interest rate gives one point on the demand curve. As the interest rate changes, the amount our hypothetical investor wishes to lend also changes. In fact, if the interest rate is low enough, the investor may change from a lender to a borrower. By

varying the interest rate, the supply and demand curve can be traced out, and the equilibrium interest rate can be determined. The equilibrium interest rate is that rate at which the amount investors wish to borrow is equal to the amount investors wish to lend. This is often called a *market clearing condition*. The equilibrium interest rate depends on what each investor's decision problem looks like, or the characteristics of a figure like Figure 1.3 for each investor. Figure 1.3 depends on the investor's income in the two periods and the investor's tastes or preferences. Thus, in this simple world, equilibrium interest rates are also determined by the same influences: investors' tastes and investors' income.

CONCLUSION

This simple example has revealed the elements that are necessary to analyze a portfolio problem. We need two components to reach a solution: a representation of the choices available to the investor, called the opportunity set, and a representation of the investor's tastes or preferences, called indifference or utility curves. With these two components we solved this simple problem and can solve the more realistic problems that follow. In addition, this simple example taught us that by aggregating across investors, we can construct models of equilibrium conditions in the capital markets. Now we turn to an examination of why and how this framework must be modified to deal realistically with multiple investment alternatives.

MULTIPLE ASSETS AND RISK

If everyone knew with certainty the returns on all assets, then the framework just presented could easily be extended to multiple assets. If a second asset existed that yielded 10%, then the opportunity set involving investment in this asset would be the line $A'BC'$ shown in Figure 1.4. Its intercept on the vertical axis would be $10{,}000 + (1.10)(10{,}000) = \$21{,}000$, and the slope would be $-(1.10)$. If such an asset existed, the investor would surely prefer

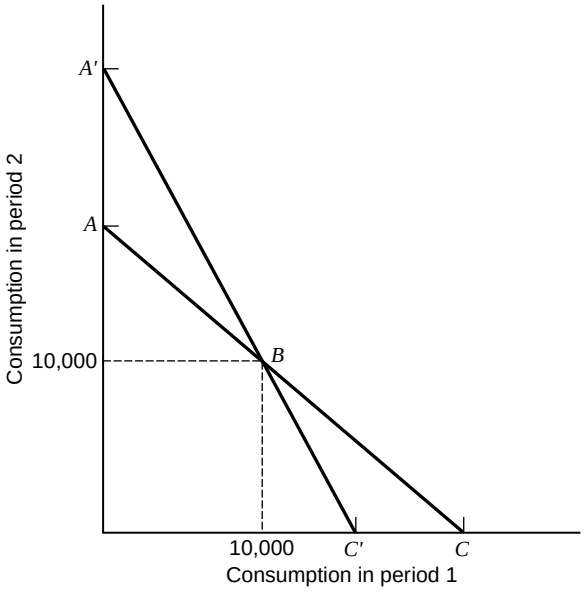


Figure 1.4 Investor's opportunity set with several alternatives.

it if lending and prefer the 5% asset if borrowing. The preferred opportunity set would be A', B, C. Additional assets could be added in a straightforward manner. But this situation is inherently unstable. Two assets yielding different certain returns cannot both be available because everyone will want to invest in the higher-yielding one and no one will purchase the lower-yielding one. We are left with two possibilities: either there is only one interest rate available in the marketplace or returns are not certain.[1] Because we observe many different interest rates, uncertainty must play an important role in the determination of market rates of return. To deal with uncertainty, we need to develop a more complex opportunity set.

The remainder of this book is concerned with the development of the framework necessary to solve the more complex asset choice problems in the presence of risk. In the next two chapters we deal with the basic notions of the investor's opportunity set under risk.

QUESTIONS AND PROBLEMS

1. Walking down an unfamiliar street one day, you come across an old-fashioned candy store. They have red hots five for one penny, and rock candy—one small piece for one penny. You decide to purchase some for yourself and your friends, but you find that you have only $1.00 in your pocket. Construct your opportunity set both geometrically and algebraically. Draw in your indifference map (set of indifference curves). Explain why you have drawn your indifference curves as you have drawn them.

2. Let us solve a two-period consumption investment decision similar to the one presented in the text. Assume that you have income equal to $20 in each of two periods. Furthermore, you have the ability to both lend and borrow money at a 10% rate. Draw the opportunity set and your indifference map. Show the optimum amount of consumption in each period.

3. Assume you can lend and borrow at 10% and have $5,000 in income in each of two periods. What is your opportunity set?

4. Assume you can lend and borrow at 5% and have $20,000 in income in each of two periods. Further assume you have current wealth of $50,000. What is your opportunity set?

5. An individual has two employment opportunities involving the same work conditions but different incomes. Job 1 yields $Y_1 = 50$, $Y_2 = 30$. Job 2 yields $Y_1 = 40$, $Y_2 = 40$. Given that markets are perfect and bonds yield 5%, which should be selected?

6. Assume you have income of $5,000 in each of two periods and can lend at 10% but pay 20% on borrowing. What is your opportunity set?

7. Assume your preference function P is $P = C_1 + C_2 + C_1C_2$. Plot the location of all points with $P = 50$, $P = 100$.

8. In Problem 3, what is the preferred choice if the preference function discussed in Problem 7 holds?

9. Suppose you have $10.00 to spend on dinner. There are two possibilities: pizza at $2.00 a slice or hamburgers at $2.50 a piece. Construct an opportunity set algebraically and graphically. Add indifference curves according to your own individual taste.

[1]Transaction costs, or alternative tax treatment of income from different securities, can explain the existence of some differential rates but nothing like the variety and magnitude of differentials found in the marketplace.

10. Using the two-period consumption model, solve the following problem. Assume you can lend and borrow at 5% and your income is $50 in each period. Derive the opportunity set and add your indifference curves.

11. Assume you earn $10,000 in periods 1 and 2. Also, you inherit $10,000 in period 2. If the borrowing/lending rate is 20%, what is the opportunity set? What is the maximum that can be consumed in the first period? In the second period?

12. Assume the borrowing rate is 10% and the lending rate is 5%. Also assume your income is $100 in each period. What is the maximum you can consume in each period? What is the opportunity set?

BIBLIOGRAPHY

1. Hirshleifer, Jack. *Investment, Interest, and Capital*. (Englewood Cliffs, NJ Prentice Hall, 1969).
2. Markowitz, Harry. *Portfolio Selection: Efficient Diversification of Investments.* (New York: John Wiley, 1959).
3. Sharpe, William. *Portfolio Theory and Capital Markets.* (New York: McGraw-Hill, 1970).

2

Financial Securities

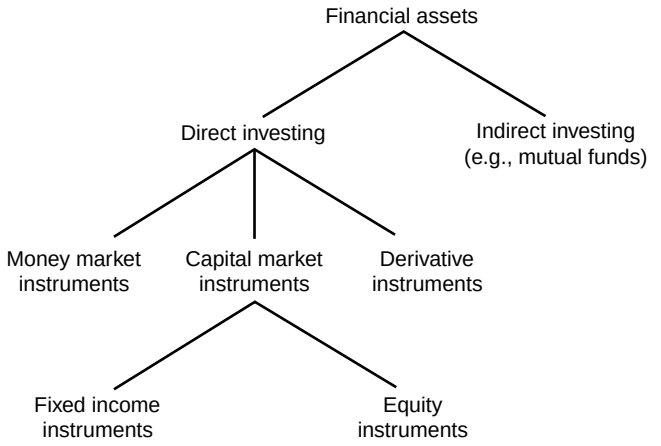
This chapter is meant to introduce the reader to the principal financial instruments, their return characteristics, and the indexes that are used to represent their returns. The nature of the material means that this chapter is much more descriptive than subsequent chapters. Those readers already familiar with financial instruments and the indexes that can be used to represent their returns can skip to Chapter 3. Those readers who have had a prior finance course and are familiar with financial instruments but are not familiar with the principal indexes used to represent their returns can skip to the section in this chapter titled "The Return Characteristics of Alternative Security Types." We can think of a *security* as a legal contract representing the right to receive future benefits under a stated set of conditions. There are a large number of financial securities. When you take out a mortgage on a house or lease a car, the contract you sign is a financial security. We are going to limit the set of financial securities we deal with by selecting primarily from among those that are traded in organized markets. In fact, Chapter 3 will focus on the nature of alternative market structures for the securities described in this chapter.

In the first section of this chapter, we describe the characteristics of a broad sample of financial securities. In the second section, we examine the performance of a representative sample of financial assets to begin to understand the relevant characteristics of different types of securities. Finally, we discuss indexes that are used to represent the performance of classes of securities. The latter material is included because in later chapters, we will often discuss market performance. We need an indication of performance and use one or more of the indexes described in this chapter.

TYPES OF MARKETABLE FINANCIAL SECURITIES

There are many ways to categorize financial securities. We have found it useful to use the scheme shown in the following diagram.

An investor can choose to purchase directly any one of a number of different securities, many of which represent a type of claim on a private or government entity. Alternatively, an investor can invest in an intermediary (mutual fund), which bundles a set of direct investments and then sells shares in the portfolio of financial instruments it holds. Because indirect investing involves purchasing shares of bundled direct investments, we discuss



indirect investing at the end of this section. Direct investment can be classified by the time horizon of the investment. Investments in debt that have a life of less than one year are usually called *money market instruments*. These can be further divided according to whether the money market instrument is issued by a government entity or a private entity. Investments with maturities of more than one year are generally called *capital market instruments*. The latter can be divided according to whether they are debt or equity instruments, and debt instruments can be further divided according to whether they are issued by a government entity or a private entity. The final category of financial assets we discuss is *derivative instruments,* so called because their payoff depends on (is derived from) the price of one of the primary assets already discussed. We now discuss each of these categories of financial assets in turn.

Money Market Securities

Money market securities are short-term debt instruments sold by governments, financial institutions, and corporations. The important characteristic of these securities is that they have maturities at the time of issuance one year or less. The minimum size of a transaction in a money market instrument is typically large, usually exceeding $100,000. In addition, some market securities that we describe are not actively traded on exchanges. Given the minimum transaction size and the inactive trading of some securities, many individuals who wish to obtain these instruments will do so by holding a mutual fund (money market fund). These funds are discussed later in this chapter. The major money market instruments are listed in Table 2.1.

Table 2.1 Money Market Instruments

Treasury bills
Repurchase agreement (repos or RPs)
LIBOR
Negotiable certificate of deposit (CDs)
Bankers' acceptances
Commercial paper
Eurodollars

In the years leading up to the financial crisis of 2008, money market securities of all kinds were increasingly used by financial institutions as relatively safe "collateral" for transactions. The financial crisis caused a flight to the safest of these, Treasury bills.

We do not intend to discuss each security in Table 2.1 in detail. We discuss three securities that play a large role in later analyses in this book and briefly summarize some general characteristics of the remaining securities.

Treasury Bills U.S. Treasury bills are the least risky and the most marketable of all money market instruments. They represent a short-term IOU of the U.S. federal government. Whereas most money market instruments are sold in minimum denominations of $100,000, Treasury bills (T-bills) are sold in minimum denominations of $10,000. New T-bills are issued by the federal government at frequent intervals. New 91- and 182-day T-bills are issued weekly, whereas 52-week T-bills are issued monthly. An active secondary market with very low transaction costs exists for trading T-bills. T-bills are sold at a discount from *face value* (the cash payment the investor will receive at maturity) and pay no explicit interest payments. The difference between the purchase price and the face value constitutes the return the investor receives.[1]

Treasury bills play a special role in financial theory. Because they are considered to have no risk of default, have very short-term maturities, have a known return, and are traded in active markets, they are the closest approximations available to a riskless investment. The rate on 30-day Treasury bills will be used throughout the book to approximate the monthly riskless rate of interest.

Repurchase Agreements (Repos) A repurchase agreement is an agreement between a borrower and a lender to sell and repurchase a U.S. government security. A borrower, usually a government securities dealer, will institute the repo by contracting to sell securities to a lender at a particular price and simultaneously contracting to buy back the government securities at a future date at a specified price. The difference between the two prices represents the return to the lender.

The repo agreement is a short-term collateralized loan for which the amount of required collateral depends on the risk of the collateral. During the financial crisis, repo counterparties took large "haircuts" to the face value of repo'd mortgage-backed securities.

The maturity of a repo is usually very short (less than 14 days), with overnight repos being fairly common. Longer repos, often labeled "term repos," may have maturities of 30 days or more. The institution on the opposite side of the repo is said to have a reverse repo. The party doing the reverse repo contracts to buy a security at a particular price and to sell it back at a predetermined price and time.

Repos and reverse repos play an important role in the pricing of derivative securities because they allow short positions to be taken in bonds. The ability to use repos is important in the type of arbitrage arguments made in future chapters.

Other Short-Term Instruments Although all short-term instruments are considered to have very low risk, they do tend to offer slightly different returns according to the type of, and even the specific, institution that offers them. CDs (negotiable certificates of deposit) are time deposits with a bank. Bankers' acceptances are contracts by a bank to pay a specific sum of money on a particular date. Both instruments sell at rates that depend on the credit rating of the bank that backs them, although CDs are insured by the Federal Deposit Insurance Corporation up to a limit of $10,000. Eurodollar and Eurodollar CDs are dollar-denominated deposits

[1]The pricing conventions on T-bills are described in Chapter 21, on bond pricing.

backed by a foreign bank or a European branch of an American bank. Because foreign banks are often subject to less regulation than U.S. banks, instruments issued by foreign banks usually carry higher interest payments than similar instruments issued by U.S. banks. Commercial paper is a short-term debt instrument issued by large, well-known corporations, and rates are determined in part by the creditworthiness of the corporation.

In a later section of this chapter, we use data on one-month Treasury bills to represent the behavior of money market instruments. Although this will serve as an example, keep in mind that other money market instruments will offer different returns because of both differences in maturity and differences in the risk of the issuing institutions. For example, when oil prices dropped dramatically and Texas real estate prices quickly followed, the creditworthiness of Texas banks declined, and CDs in Texas banks sold at much higher yields than did average CDs.

Before leaving this section, we discuss an important element of money markets that is not an instrument but rather a rate.

The London Interbank Offered Rate (LIBOR) LIBOR is the rate at which large international banks in London lend money among themselves. We single it out for special mention because it is used as a base rate for many types of longer-term loans, even in U.S. markets. Despite the fact that it is a rate between London banks, it is usually quoted for loans in dollars. It is quite common to see longer-term debt instruments with rates that change periodically (and thus with some of the characteristics of shorter-term instruments). These changing rates are usually set at either the Treasury bill rate plus a fixed amount or at the LIBOR rate plus a fixed amount.

Capital Market Securities

Capital market securities include instruments with maturities greater than one year and those with no designated maturity at all. The market is generally divided according to whether the instruments contain a promised set of cash flows over time or offer participation in the future profitability of a company. The first sector is usually referred to as the fixed income market, whereas the second is the equity market. Preferred stock, discussed last, is an instrument that has some of the characteristics of each of the other two types.

Fixed Income Securities Fixed income securities have a specified payment schedule. Most are traditional bonds and promise to pay specific amounts at specific times.[2] Usually this is in the form of prespecified dates for the payment of interest and a specific date for the repayment of principal. In almost all cases, failure to meet any specific payment puts the bond into default, with all remaining payments (missed interest plus principal) due immediately. Fixed income securities vary in promised return because of differences including the maturity of the bond, the creditworthiness of the issuer, and the taxable status of the bond. We will start by examining the safest type of fixed income securities, those offered by the U.S. government.[3]

Treasury Notes and Bonds The federal government issues fixed income securities over a broad range of the maturity spectrum. Debt instruments from 1 to 10 years in maturity are called *Treasury notes.* Debt instruments with a maturity beyond 10 years are known as

[2]Many of the capital market securities are discussed in much more detail in later chapters of this book. For example, the pricing and management of fixed income securities is discussed in more detail in Chapters 20 and 21.

[3]One should be aware that the quoted price of a fixed income security is not what the investor pays to purchase the security; rather, the investor pays the quoted price plus interest accrued since the last coupon payment.

Treasury bonds. Both notes and bonds pay interest twice a year and repay principal on the maturity date. One difference between Treasury bonds and notes is that some bonds are callable before maturity (most often during the last five years of the bond's life), while notes are not callable.[4] *Callability* means that the government can force the holder of the bond to sell the bond back to the government according to a fixed schedule of prices before maturity. For example, if a bond is callable at $101, the government has the option of buying the bond back at $101. The government would likely exercise the option when it benefits itself, and thus this is disadvantageous to the investor. Thus callable bonds have to offer the investor a higher return to compensate for the possibility of a disadvantageous call.

Treasury instruments are generally considered to be safe from default, and thus differences in expected returns are due to differences in maturity, differences in liquidity, and the presence or absence of a call provision.[5]

Federal Agency Securities Federal agency securities are issued by various federal agencies that have been granted the power to issue debt to help certain sectors of the economy. For example, the Farm Credit Banks make funds available for such things as research and short-term loans to farm cooperatives. Federal agency securities are often thought of as a close substitute for Treasury securities. Although federal agency securities are not backed by the full faith and credit of the federal government, investors assume that the federal government would not allow an agency to default in its payments. However, the lack of an explicit guarantee from the federal government plus the fact that markets for agencies are frequently less liquid than markets for Treasury instruments has resulted in the agency instruments selling at slightly higher yields than Treasury notes and bonds.

Municipal Bonds Municipal bonds are debt instruments sold by political entities, such as states, counties, cities, airport authorities, school districts, and so forth, other than the federal government or its agencies. They differ from agency bonds in that they can (and in rare instances do) default and their interest is exempt from federal and usually (within the state that issues them) state taxes. The principal types of municipal bonds are general obligation bonds, which are backed by the full faith and credit (taxing power) of the issuer, and revenue bonds, which are backed either by the revenues of a particular project (e.g., a toll road) or the particular municipal agency operating the project.

Because of the tax-exempt feature of municipal bonds, they sell at lower promised yields than nonmunicipal bonds of the same risk. To find an equivalent yield, one must explicitly compare the discounted value of after-tax cash flows with before-tax cash flows. It is common practice to use the following approximation to the taxable equivalent yield:

$$\text{Taxable equivalent yield} = \frac{\text{Tax-exempt municipal yield}}{1 - \text{Marginal tax rate}}$$

This approximation holds exactly only if municipal bonds sell at par, the treasuries they are being compared to sell at par, and the yield curve is flat. One must be particularly careful using this approximation for municipal bonds selling below par: while the interest payment on municipal bonds is tax exempt, capital gains are subject to taxation.

Corporate Bonds Corporate bonds are generally similar to government bonds in payment pattern. They promise to pay interest at periodic intervals and to return principal at a

[4]Treasury bonds issued after 1985 do not contain call provisions.

[5]The tax implications of different coupon rates can also explain differences in yield; this will be discussed in later chapters.

fixed date. The major difference is that these bonds are issued by business entities and thus have a risk of default. Corporate bonds are rated as to quality by several agencies, the best known of which are Standard and Poor's and Moody's.[6]

Corporate bonds differ in risk not only because of differences in the probability of default of the issuing corporations but also because of differences in the nature of their claims on the assets and earnings of the issuing corporations. For example, secured bonds have specific collateral backing them in the event of bankruptcy, whereas unsecured corporate bonds (called *debentures*) do not. An additional class of bonds called *subordinated debentures* not only have no specific collateral but also have a still lower priority claim on assets in the event of default than unsubordinated debentures. In an attempt to gain some protection against bankruptcy, corporate bonds typically place certain restrictions on management behavior as part of the loan agreement (called the *bond indenture*). Such restrictions might include limiting the payment of dividends or the addition of new debt.

Another notable feature of corporate bonds is that they are most often callable, which means that corporations can force the holder of the bond to surrender them at a fixed price (usually above the price at which the bonds were initially sold) during a set period of time. Corporations usually call bonds at a time when interest rates are below those that existed when the bond was first sold. Thus the bondholder risks reinvesting his or her proceeds from a call at lower rates than the interest rate of the bond at the time of issuance.

Not-So-Fixed Income Securities It is evident from the preceding discussion that fixed income securities do not always pay the security holder the promised payment (because of calls or default). This leads to variability in cash flows received by the investor. Two classes of fixed income securities have even greater variability in cash flows: preferred stocks and mortgage-backed securities. In both cases, variability in cash flows is expected, and variability does not result in the holder's right to force bankruptcy.

Preferred Stock Preferred stock at first blush resembles an infinite life bond. It promises to pay to the holder periodic payments like coupons, but these are called dividends rather than interest. There is no return of principal in this case because preferred stock is almost always infinite in life. Preferred stock is not really a fixed payment instrument, however, in that failure to pay the promised dividend does not result in bankruptcy. Usually when a firm fails to pay dividends, these dividends are cumulated, and all unpaid preferred stock dividends must be paid off before any common stock dividends can be paid.

Preferred stock occupies a middle position between bonds and common stock in terms of priority of payment of income and in terms of return on capital if the corporation is liquidated. In addition, most preferred stock does not actually have an infinite life because the issues are frequently callable, and many of the issues may be converted into common stock at the discretion of the holder. Of course, a combination of callability and convertibility allows the issuer to force conversion. These features affect the risk and reward from holding preferred stocks.

Asset-Backed Securities An asset-backed security is a contractual claim on a pool of securities—typically loans. These include home mortgages, commercial mortgages, automobile loans, student loans, and credit card debt. Collectively referred to as collateralized debt obligations (CDO), they are usually structured so that there are several classes, known as *tranches*, with different maturities and different levels of risk. Collateralized mortgage obligations are backed by pools of mortgages, and CDOs are

[6]See Chapter 21, on bond pricing, for a more detailed description of bond ratings and their impact on bond prices.

backed by pools of commercial or personal loans. Collateralized bond obligations are backed by low investment-grade corporate bonds. Asset-backed securities are the product of a series of financial innovations in the late twentieth century. They also played a central role in the financial crisis of 2007–2008 because they represented a large component of the portfolio of most global financial institutions. Mortgage-backed securities were particularly important in this regard.

Mortgage-Backed Securities The last "not-so-fixed income" security type that is most often classified as a fixed income security is the mortgage-backed security, which represents a share in a pool of mortgages. The best-known mortgage-backed security is the Ginnie Mae (GNMA), which are issues of the Government National Mortgage Association. These instruments are backed by the full faith and credit of the U.S. government, so the investor bears no default risk. However, the investor is subject to considerable interest rate risk. These instruments are "pass-through" securities, which means that all interest and principal payments on the individual mortgages making up the pool backing a particular GNMA certificate are paid (passed through) to the holder of a GNMA. The stated maturity in GNMAs may be as high as 40 years, but the average life is considerably shorter. The pass-through feature means that the holder will receive a very uncertain stream of future payments because it is dependent on how fast mortgage holders pay off their mortgages. Furthermore, to the extent that mortgages are paid off when interest rates are low, the investor receives funds at the time when investment opportunities have expected returns below the promised return on the original GNMA. The added element of risk is compensated for by GNMAs selling at a higher promised return than government securities of similar expected life.

Mortgage-backed securities are also offered by several other quasi-government agencies, including Fannie Mae and Freddie Mac. Although funded by public equity issuance, these institutions were rescued by government bail-out during the financial crisis. Prior to 2008, an increasing proportion of mortgage-backed securities were issued by private institutions without explicit or implicit government guarantee. "Private-label" mortgage-backed securities funded a large share of the "subprime" loans that suffered severe price declines during the crisis. These securities carry an additional risk because they may be backed by the credit of the issuing body or simply by the pool of mortgages themselves, which are held in special-purpose vehicles (SPV), created solely to issue asset-backed securities.

Common Stock (Equity) Common stock represents an ownership claim on the earnings and assets of a corporation. After holders of debt claims are paid, the management of the company can either pay out the remaining earnings to stockholders in the form of dividends or reinvest part or all of the earnings in the business.

The unique feature of common stock (unlike simply owning the business) is that the holder of common stock has limited liability. If a company goes bankrupt, all that the holder of common stock can lose is his or her original investment in the stock. The creditor cannot look to the general assets of the stockholder to finance his claims.

Despite limited liability, because of the residual nature of its claim to earnings and assets, common stock as a class is the riskiest of the securities discussed to this point.[7]

Derivative Instruments

Derivative instruments are securities whose value derives from the value of an underlying security or basket of securities. The instruments are also known as contingent claims because

[7]Common stock issued by some companies can be less risky than some high-risk debt issues.

their values are contingent on the performance of underlying assets. The most common contingent claims are options and futures. An option on a security gives the holder the *right* to either buy (a call option) or sell (a put option) a particular asset or bundle of assets at a future date or during a particular period of time for a specified price. The buyer pays a price for this option but is free not to exercise this option if prices move in the wrong direction. A future is the *obligation* to buy a particular security or bundle of securities at a particular time for a stated price. A future is simply a delayed purchase of a security. Futures and options are securities that represent side bets on the performance of individual or bundles of securities. There is always a buyer and a seller of an option or a future, and the profit (or loss) to the seller is exactly equal to the loss (or profit) of the buyer. The action of the buyer or seller of options or futures does not affect the cash flows to the corporation, nor does it result in a change in the number or type of securities the corporation has outstanding. Another kind of side bet is referred to as a credit default swap (CDS). These are insurance contracts to protect lenders against credit defaults. Essentially, the lender pays an insurance premium to the issuer of the CDS, who will purchase the asset in the event of a default.

The corporation can issue contingent claims, however, and in this case the value of the corporation is often impacted by the action of holders of its contingent claims. Corporate-issued contingent claims include rights and warrants, which allow the holder to purchase common stocks from the corporation at a set price for a particular period of time, and convertible securities (bonds and preferred stocks), which allow the holder to convert an instrument into common stock under specified conditions. Although these corporate contingent claims have many features in common with other derivative instruments, they differ in that if the holders execute them, it results in a change in the attributes of the corporation (e.g., the receipt of cash and/or change in the nature and size of capital). This means that these contingent claims are more difficult to analyze than those not issued by the corporation.

Indirect Investing

While an investor can purchase any of the instruments described here (and several we have not touched on), the investor can instead choose to invest indirectly by purchasing the shares of investment companies (mutual funds). A mutual fund holds a portfolio of securities, usually in line with a stated policy and objective. Mutual funds exist that hold only a small set of securities (e.g., short-term tax-free securities or stocks in a particular industry or sector) or broad classes of securities (such as stocks from major stock exchanges around the world or a broad representation of American stocks and bonds).

Mutual funds come in two flavors: open-end funds and closed-end funds. Open-end fund shares are purchased (and sold) directly from (and to) the mutual fund. They are purchased (and sold) at the value of the net assets standing behind each share, where the net asset value is determined once a day, at a stated time. As a first approximation, if you own 1/100 of the shares outstanding in a mutual fund, your shares are worth 1/100 of the market value of the total portfolio of securities that the fund owns. The reason we use the term *first approximation* is that some mutual funds charge a fee when the investor buys a fund (front-end load) and some charge additional fees (back-end load) when an investor sells shares in a fund. For example, in the case of an 8% front-end load, you purchase assets with only 92% of the money you put up. Similarly, in the case of a 6% back-end load, you will receive only 94% of the value of the assets your shares represent when you sell the fund. Very often, back-end loads decrease as a function of the amount of time the investor holds the fund.

Closed-end funds differ from open-end funds in that they initially sell a predetermined number of shares in the fund. They then take the proceeds (minus costs) from the sale of fund shares and invest in stocks or stocks and bonds. Shares in the fund are then traded on an

exchange and take on a life of their own. Owning a share in a closed-end fund is like owning a share in any corporation, but the assets of the corporation are stocks and bonds.[8] Unlike open-end funds, the shares of a closed-end mutual fund can sell at a discount or a premium to their net asset value. Premiums and discounts are related to the perceived quality of management and certain tax liabilities. In fact, most closed-end funds sell at a discount from net asset value. The clear exceptions are funds such as the Korean funds, where the only way an American investor can own stocks in Korea is through buying a closed-end fund. These funds sell at a monopoly premium.

Mutual funds may offer the investor special services such as check-writing privileges or the ability to switch between mutual funds (types of investment) in the same family of funds at no cost. Although most offer liquidity, diversification, and "professional management," they do not offer these qualities without a cost. Investors pay a pro rata share of the expenses and management fees charged by the mutual fund company. In addition (for "open-end" funds), investors may pay a sales charge and/or a special charge known as a 12b-1 fee, which is a fee charged to the customer of a fund to compensate the fund for the cost of promoting (e.g., advertising) the fund. We will examine additional attributes of mutual funds in Chapter 25.

THE RETURN CHARACTERISTICS OF ALTERNATIVE SECURITY TYPES

When describing securities in the previous section, we alluded to risk and return. One of the basic tenets of this book is that investors like high return but don't like high risk. Although we will be much more specific about measuring risk and return in future chapters, it is useful to become familiar with the risk and return characteristics of some of the securities we have discussed.

First, we should discuss what we mean by *return*. We will in most instances use *return* to indicate the return on an investment over a particular span of time called *holding period return.* Return will be measured by the sum of the change in the market price of a security plus any income received over a holding period divided by the price of a security at the beginning of the holding period. Thus, if a stock started the year at $100, paid $5 in dividends at the end of the year, and had a price of $105 at the end of the year, the return would be 10%.[9]

In describing securities, we mentioned several factors that should affect risk. These included

1. the maturity of an instrument (in general, the longer the maturity, the more risky it is)
2. the risk characteristic and creditworthiness of the issuer or guarantor of the investment
3. the nature and priority of the claims the investment has on income and assets
4. the liquidity of the instrument and the type of market in which it is traded.[10]

If risk is related to these elements, then measures of risk such as the variability of returns should be related to these same factors.

Although we will not introduce formal measures of risk until Chapter 4, let us just state at this time that a widely accepted measure of risk (metric for capturing historic variability)

[8]There is a difference in that income, if paid out to fund shareholders, is not subject to corporate taxes provided certain conditions are met by the fund.

[9]If dividends or other income is received during the period of time over which return is computed, an assumption must be made about the rate at which the cash flow is reinvested until the end of the period.

[10]This will be discussed in Chapter 3.

Table 2.2 Return and Risk for Selected Types of Securities in Percent per Year (1926–2011)

	Average Return	Standard Deviation
U.S. Treasury bills	3.6	3.1
Long-term government bonds	6.1	9.8
Large company stocks	11.8	20.3
Inflation	3.1	4.2
Long-term corporate bonds	6.4	3.4
Small company stocks	16.5	32.5

Source: Ibbotson and Associates, 2008.

is called the standard deviation. The standard deviations of the return series of several instruments, together with the average return for each series, are shown in Table 2.2.

One of the major tenets of this book is that returns (over long periods of time) should be consistent with risk. In fact, the average historical returns presented earlier are broadly consistent with this.

An examination of historical returns on security types such as those presented earlier is frequently used as a starting point for preparing forecasts of the return expected from broad classes of assets.

For example, a forecaster might start with a forecast of the inflation rate over the next year. Economists argue that Treasury bills over long periods of time should compensate investors for any loss in purchasing power (inflation) plus the time value of money (giving up the use of funds for a short period of time). From Table 2.2 we see that the return on Treasury bills has averaged 0.5% per year over inflation. One forecast of the return for Treasury bills would simply be to add 0.5% to the forecast of inflation. Alternatively, one could simply use current rates.

In a similar manner we can note that the historic difference in returns between Treasury bills and long-term Treasury bonds is 2.5%. However, as with all historical premiums, the historical premium might be modified to reflect current beliefs about the future, relative to the past. However, historical data provide a useful starting place. This is often called a *term premium*. Although it depends on supply and demand conditions to the capital markets and the pattern of longer-term expectations about the movement in short-term rates, the term premium also depends on the risk preferences of investors. Longer-term bonds have more variable returns than short-term bonds.

The difference between long-term corporate bond rates and the rate on long-term government bonds (0.3%) is compensation in part for the greater risk of default of corporate bonds.

The rate of return on large common stocks is $11.8 - 6.4 = 5.4\%$ higher than the rate of return on long-term corporate bonds because of the greater risk associated with the future cash flows on large stocks. The rate on small stocks is $16.5 - 11.8 = 4.7\%$ higher than the rate on large stocks, due in part to the added risk associated with small stocks.

The type of building-block approach to rates of return on security types presented earlier is frequently used to forecast rates of return in the future. That is, starting with either forecasts of inflation or the Treasury bill rate, management will modify historic differentials to estimate the expected returns on categories of securities. Modifications of the past differentials for forecasting are based on forecasts of supply and demand conditions in different capital markets as well as on forecasts of general economic activity.

We have indicated the return performance of some broad classes of securities. In doing so, we have used a set of performance indexes without actually describing the indexes. Because we will often talk about indexes of general performance and "the market" in this book, it is worthwhile spending a short amount of time reviewing some widely used indexes.

STOCK MARKET INDEXES

The oldest continuously quoted index of stock price performance in the United States is the Dow Jones Industrial Average Index (DJIA); this index has been computed since 1896. Since 1928, it has consisted of a price-weighted average of 30 large "blue chip" stocks. When the index was originally constructed, it contained 20 stocks, and the value of the index was found by adding the prices of the 20 stocks (assuming the investor bought one share of each stock). Today the average is computed by adding the prices of 30 stocks and dividing by an adjustment factor.[11] Despite the fact that this index is the most widely quoted stock market index, it has some flaws. First of all, 30 stocks, particularly 30 stocks that are among the largest, represent at best a very narrow definition of the market.

Second, and perhaps most important, the implicit price weighting in the index assumes that an investor is equally likely to buy one share of any stock. Another way to view this is that the investor is more likely to place a dollar in a share of stock if it sells at a higher price. The "market" represents the aggregate of the action of all investors. All investors in aggregate must hold all stocks in proportion to the fraction that the aggregate market value of any stock represents of the total market value of all stocks. This is clearly different than the DJIA, which weighs each stock by the price of that stock relative to the sum of the market value of one share of each stock in the index. The absurdity to which this weighting can lead is evident by what happens if a stock splits. In a two-for-one stock split, the weighting of the split stock after the split would be one-half of the weighting before the split.

Despite these defects in the methodology used in computing this index, the DJIA continues to be widely employed and mimicked. For example, one of the most widely used indexes of the Japanese stock market, the Nikkei 225, is computed in the same manner as the DJIA. The index does allow the rate of price increase to be computed for a well-defined strategy: buy one share of each stock in the index, selling off any additional shares received due to stock splits or stock dividends, while reallocating the proceeds among all shares in the index. To compute the rate of price appreciation from this strategy, one simply takes the change in the index over a certain period of time and divides by the value of the index at the beginning of the period. Note that this provides only a rate of price appreciation, not a total rate of return, for dividends are ignored in computing the index.

Most stock price indexes are weighted by market capitalization. The next most popular index of the U.S. stock market is the Standard and Poor's Composite 500 stock index (S&P 500). In calculating this index, the price of each of the 500 stocks is multiplied by the market value of the company's shares outstanding, divided by the aggregate market value of all 500 companies. We can think of this index as reporting the price performance of a portfolio where the investor buys the same percentage of the total outstanding stock (in market value) of each company. Note that stock splits and dividends do not affect the index because they have no effect on the total market value of the outstanding stock.[12]

The reader should note that the Standard and Poor's index does not include dividends; thus using it directly allows the computation of a rate of price appreciation and not a rate of return. A crude adjustment for dividends (to get total return) can be achieved by splicing the Standard and Poor's price index with the dividend yield index published by Standard and Poor. In recent years, however, a number of sources, most notably the Center for Research

[11]The adjustment factor is computed so as to correct for discontinuities in the index caused when a stock is substituted for one previously in the index or when a stock in the index has a stock split or dividend.

[12]The stock split could affect expectations about the future cash flows of the firm and its market value.

in Security Prices (CRSP), have computed a version of the S&P index corrected for dividends. This is the index we used in the previous section to represent common stocks (large stocks).[13]

In recent years the number of indexes measuring common stock performance here and abroad has proliferated. Large populations of stocks are represented in the United States by, among others, the New York Stock Exchange (NYSE) index (including all stocks listed on NYSE), the Amex index, the Wilshire 5000 stock index (NYSE, American Stock Exchange, actively traded over-the-counter stocks), and so forth. They are all market-weighted indexes, though they do not include dividends. CRSP has available a number of return indexes for different groups of stocks on the New York, American, and over-the-counter markets. These are calculated on a market-weighted basis and include return from reinvestment of dividends.

Finally, a number of international stock market indexes are market weighted and are computed with dividends. For example, Morgan Stanley International computes indexes for more than 20 countries as well as for different geographical sectors of the world and an Aggregate World index.

BOND MARKET INDEXES

Although almost all of the major stock market indexes exclude dividends and thus are not total return indexes, the major bond indexes are total return indexes, for they include interest payments as well as capital gains. The best-known bond indexes are constructed by Barclays, FTSE, and MSCI (Morgan Stanley Capital International).[14] They are all market-weighted total return indexes including all issues above a certain size. Furthermore, subindexes exist covering different parts of the bond market by maturity as well as by type of issuer.

Perhaps the use of market weighting and the inclusion of cash flows (interest) in the indexes reflect the fact that bond indexes were constructed more recently than stock indexes, when the concepts of market weighting and total return were better understood.

One caution on using these indexes is that a number of issues in the indexes are not actively traded. The prices of these issues represent price estimates based on issues that are traded; this estimation process can be a source of inaccuracy.

The set of bond indexes with the longest history are those compiled by Ibbotson and Associates. Ibbotson reports monthly returns from the beginning of 1926 to the present for Treasury bills, long-term government bonds, intermediate-term government bonds, and long-term corporate bonds. These series are excellent for gaining perspective on the major bond markets because of their long history. However, the user of these series should be aware that the numbers of bonds included in each of the series are not the same. For example, (1) the long-term corporate bond series currently includes nearly all Aaa- and Aa-rated corporate bonds, whereas (2) the long-term government bond series is based on a single government bond of approximately 20 years' maturity selected at the start of each year.[15]

---

[13]The monthly version of the CRSP index treats all dividends paid on a stock as reinvested at the end of the month. The assumption is implicitly made that cash payments earn no return during the remainder of the month in which they are paid.

[14]These indexes differ in both the population of bonds included and the assumption about the timing of the reinvestment of interest payments.

[15]See Ibbotson annual yearbooks for a detailed description of the construction of each of the bond series.

CONCLUSION

We have described the attributes of a broad representation of financial assets in this chapter. We have looked at some indexes that are used to measure the performance of broad classes of assets, and we have examined in risk and return terms the characteristics of a representative set of assets. We have not discussed the markets these assets trade on nor the impact of market structure on the characteristics of assets; these will be discussed in Chapter 3.

3

Financial Markets

Almost every chapter in this book is concerned with selecting securities, constructing portfolios, and evaluating these decisions. In this chapter we will discuss how securities are traded and the nature of the markets in which they are traded.[1] This chapter, like Chapter 2, is more descriptive and less analytical than the rest of the book. The reader who is familiar with the mechanics of the markets in which securities are traded or who is not concerned with this subject can go directly to Chapter 4 with no loss of continuity.

The characteristics of markets can influence trading costs, the speed with which information is reflected in prices, and the accuracy with which prices reflect available information. Thus characteristics of markets can determine how often one should trade as well as the degree of mispricing of a security (or suboptimality of a portfolio) before a trade could be profitable.

The chapter is divided into four sections: (1) the mechanics of trading of a security, (2) margin, (3) the nature and structure of markets, and (4) special characteristics of trades, including their type and costs.

TRADING MECHANICS

An individual wishing to buy or sell a security would first establish an account with a brokerage firm and then submit (by phone or computer) an order. The order specifies

- the security the investor wishes to trade (e.g., the 2 $^{3}/_{4}$ % coupon U.S. Treasury bond maturing on February 15, 2019, or General Motors common stock)
- the direction of the order (buy or sell)
- the order size (e.g., five $10,000-par bonds or 400 shares of stock)
- whether the order is market or limit (and related qualifications)

The last is discussed in more detail.

[1]The authors wish to thank Professor Joel Hasbrouck for his excellent advice in revising this chapter.

Market Orders A market order directs the broker to buy or sell the security at the best available price. For example, assume IBM is quoted $80.20 bid and offered at $80.30 (per share). This means that someone is willing to buy at $80.20, and someone else (or perhaps the same trader) is willing to sell at $80.30. Thus an investor placing a market order to buy 100 shares would expect to pay $80.30 per share (a total of $8,030) plus the commission (the broker's fee). An investor placing a market order to sell 100 shares would expect to receive $80.20 per share (a total of $8,020) less the commission. The investor might actually pay or receive a price that differs from the quote. This usually happens because the quotes have changed while the order is being conveyed to its destination. The other possibility is that in the course of conveying and executing the order, someone is willing to do the trade at a price that betters the visible quote, an outcome called *price improvement*. A broker who accepts a market order has the responsibility to execute the order as quickly as possible. The broker cannot, for example, wait to see if he or she can get the customer a better price.

The difference between the bid and ask prices is the spread. The spread can be considered a part of the overall trading cost. An investor who buys and then immediately sells using market orders, paying the ask and receiving the bid, will lose the spread (plus commissions). The most actively traded U.S. stocks generally have spreads of less than $0.05 per share, but stocks that are infrequently traded might exhibit spreads of $0.50 or more.

Limit Orders With a limit order, the customer states a price that specifies the worst acceptable terms of trades. “Buy 100 shares of IBM, limit $75.00” instructs the broker not to pay more than $75.00 per share. A purchase price that betters the limit price is acceptable, such as $74.90. Similarly, “sell 100 shares of IBM, limit $85.00” tells the broker not to accept less than $85. With a limit order, execution is uncertain. The broker will attempt to execute the order, but if the current market bid is well above $75.00, it is unlikely that an order to buy at that price will be executed. If there were to be a trade below $75.00, however, a broker that failed to execute the limit order would be said to have *missed the market* and mishandled the order. Unless otherwise specified, unexecuted limit orders are usually cancelled at the end of the day. Customers can direct, however, a *time in force* (such as 10 seconds or one hour).

Short Sale Investors can sell securities they do not own. This type of trade is referred to as a short sale. When an investor short sells a security, a security is physically sold. Because the investor does not own the security, the brokerage firm borrows it from another investor or lends the security to the investor itself. The securities borrowed normally come from the securities held at the brokerage firm for other investors. Securities kept at a brokerage firm by investors are referred to as securities registered in street name. For example, an investor might wish to short sell 100 shares of General Motors. If the brokerage firm had 100 shares of General Motors in street name, and the owner of these shares had given the brokerage firm permission to use these shares for short sales, they would sell those shares. If the firm did not possess the shares it desired to sell, it would borrow the shares from someone else, often another broker. The investor whose shares were borrowed and sold normally would not know that the transaction had occurred and would definitely not know who had borrowed the shares. Because the shares are physically sold, the company would not pay dividends to the investor whose shares were borrowed but instead would pay the purchaser of the shares. For the investor whose shares were borrowed not to be hurt by the short sale, he or she must receive the dividends. The person who sold the shares short is responsible for supplying the funds

so that the person whose shares were borrowed can receive any dividends paid on the stock that was sold short.

At a future time the short seller repurchases the shares and replaces the shares that were borrowed. Thus capital gains and losses are equal in magnitude but opposite in sign to a short seller compared to a purchaser of the shares. Because the short seller pays dividends to the person whose shares were borrowed, and the capital gains or losses to a short seller are exactly opposite those of a purchaser, the return to the short seller is minus the return of the purchaser.[2]

The textbook reason for short sales is that the short seller expects the shares to decline in value and wishes to profit from the decline. For example, assume General Motors shares are at $60 and the investor believes they will decline to $50. If the investor short sells the securities and is correct in the expectation of decline, then the investor repurchases the securities when they decline to $50 and thus makes $10 a share. There are other reasons for short sales, however. The principal one is to decrease the sensitivity of a portfolio to market movements. Securities rise and fall because of general market conditions as well as events specific to the security or a subset of securities. Because the return on short sales is the opposite of the return on a long purchase, a portfolio that includes short sales as well as long purchases reduces the exposure to market movements.[3]

Stop Orders A fourth type of order is one that is activated only when the price of the stock reaches or passes through a predetermined limit. The price that activates the trade is called a *stop price*. Once a trade takes place at the stop price, the order becomes a market order. For example, a stop loss order at $40 is activated only if trades of others take place at $40 or less. If trades take place at $40 or less, the order is activated and the order becomes a market sell order. Thus the stop loss order can be viewed as a conditional market order. A stop buy order becomes a market buy when the trades of others equal or exceed the stop price. For example, a stop buy order at $50 becomes a market order when trades take place at $50 or above.

Stop loss orders are used to attempt to lock in a gain. For example, consider an investor who purchased shares at $20 and subsequently saw the price rise to $50. The investor might place a stop loss order at $45. If the share price declines, the investor still expects to gain ($45 − $20 = $25). If the price continues to rise, then the investor continues to hold the shares and benefits from the rise. A stop loss order might be appropriate if the investor believes that the stock is overpriced and might decline but believes it is likely to rise even more before other investors reach the same conclusion. In this case the investor might place a stop loss order increasing the stop price if the shares continue to rise. As with all market orders, the actual price of the shares will trade at is uncertain because the trade prices might move below the stop price before the stop loss order can be executed.

A stop buy order is often used in conjunction with a short sale. Recall that a short sale is a sale of a security one does not own. Because the share must be replaced at a later date, a price increase harms the short seller. A stop buy order serves to limit the amount of the loss the short seller can incur.

Length of Time an Order Is Outstanding

For orders other than market orders, an investor must specify the length of time the order is to be outstanding. A day order instructs the broker to fill the order by the end of the day.

---

[2]Normally the short seller neither receives interest on the proceeds of the sale nor pays interest to the person whose securities were borrowed, although either possibility can occur.

[3]There are alternative ways to reduce market exposure. See Chapter 24.

If the order is not filled by the end of the day, the order is automatically canceled. If the investor does not specify the length of time the order is to be outstanding, it is assumed to be a day order. A week or month order instructs the broker to fill the order by the end of the week or month or cancel the order. Good-until-canceled orders remain outstanding until the investor specifically cancels the order. Finally, fill-or-kill orders instruct the broker to fill the order immediately or to kill the order.[4]

MARGIN

Investors can buy securities either with cash or with part cash and part borrowing. If the investor utilizes borrowing as well as cash, the investor is said to purchase the securities on margin. An investor utilizing margin borrows money from the brokerage firm, which in turn borrows the money from a bank. The securities purchased serve as collateral for both the brokerage firm and the bank. Thus an investor utilizing margin must leave the securities with the brokerage firm rather than take delivery (called leaving securities in "street name"). In addition, the investor signs a hypothecation agreement that allows the brokerage firm to use the customer's securities as collateral for its own loans and to lend the securities to others.

The customer is charged an interest rate on the loan. This rate is determined by adding a premium (usually 1%) to the rate the brokerage firm is charged on its loan (designated as the call rate). The amount the customer can borrow to finance a purchase or short sale is carefully regulated; these regulations are referred to as initial margin requirements. There are separate regulations that monitor the amount of the loan relative to the value of the assets at each point in time; these are called *maintenance margin requirements*. Finally, the way margin is defined for an account with long purchases is different from the way margin is defined for an account with short sales. Thus an account with both long purchases and short sales must meet both sets of margin requirements.

Margin Long Purchase

Margin for long purchases is defined as[5]

$$\text{Margin} = \frac{\text{Market value of assets} - \text{Amount borrowed}}{\text{Market value of assets}}$$

For example, if 100 shares of AT&T were purchased at $50 a share and the purchase was partially financed with a loan of $2,000, then the investor's account would look like this:

ASSETS		LIABILITIES	
100 shares of AT&T	$5,000	Loan	$2,000
		Net worth	$3,000
			$5,000

and the margin is

$$\text{Margin} = \frac{3{,}000}{5{,}000} = 60\%$$

[4]There are other types of specialized instructions that can be given, such as specifying that a market order be executed at the close.

[5]Not all securities are counted in calculating margin. For example, securities that are not readily traded, such as securitized partnerships in private deals, are not counted in determining assets.

As time passes, the margin in the account will vary as security prices change. For example, if AT&T were to increase to $70 a share, the account would be

ASSETS		LIABILITIES	
100 shares of AT&T	$7,000	Loan	$2,000
		Net worth	$5,000
			$7,000

and the margin is

$$\text{Margin} = \frac{5{,}000}{7{,}000} = 71.43\%$$

Initial Margin Long Purchase

The minimum amount of margin that must be in the account immediately after a security is purchased is called *initial margin*. The initial margin requirement is set by the board of governors of the Federal Reserve System (although an individual brokerage firm can set it higher). This requirement has varied considerably over time and has been as high as 100%, which precludes any borrowing for new purchases.

Margin is one of the tools utilized by the Federal Reserve to influence the economy. Consider an initial margin requirement of 60%. Assume the investor opens an account and purchases 100 shares of AT&T at $50 a share, or a $5,000 purchase. The investor would need 0.60 × $5,000 or $3,000 in cash and could borrow the remainder of $2,000. If the initial margin requirement was 80%, then the investor would need 0.80 × $5,000, or $4,000 in cash, with the remainder being borrowed. For accounts that already include borrowing, the amount of cash an investor needs for an additional purchase depends on the price movements of the securities owned and the amount of prior borrowing.

Consider the investor who bought 100 shares of AT&T at $50 a share, paying for the purchase with $3,000 in cash and $2,000 in borrowing. If, subsequently, AT&T were to increase in price to $70 a share and initial margin requirements were 60%, the investor could purchase 100 shares of Bethlehem Steel at $10 only utilizing borrowing, because, after the purchase, the account would look like this:

ASSETS		LIABILITIES	
100 shares of AT&T	$7,000	Loan	$3,000
100 shares Bethlehem Steel	$1,000	Net worth	$5,000
	$8,000		$8,000

and the margin would be above the initial margin requirement because

$$\text{Margin} = \frac{5{,}000}{8{,}000} = 62.50\%$$

Thus initial margin regulates the amount that can be borrowed at the time securities are purchased. The amount that can be borrowed can vary from zero to more than 100%. It could be more than 100% if the securities in the account had declined significantly in value, because at the time of any new purchase, initial margin requirements must hold for the whole account.

The securities serve as collateral for the investor's loan and for the broker's loan. To guarantee that the loans can be paid, there is a lower limit to which the margin can fall without the investor having to put up additional security. This is the subject of the next section.

Maintenance Margin Long Purchase

The minimum amount to which the margin can decline without an investor having to take action is called the maintenance margin. The maintenance margin is set by the exchanges, although an individual brokerage firm can set it higher. If the stockholder's margin drops below the maintenance margin, then the brokerage firm issues a margin call. The shareholder must bring the margin above the maintenance margin by either adding additional cash or securities to the account or selling securities. If the investor fails to respond to the margin call or the investor is unable to be reached by the brokerage firm, the brokerage firm sells off sufficient securities to bring the margin above the maintenance margin. Usually initial margin requirements are substantially higher than maintenance margin requirements. Thus there could be a substantial decline in price without a margin call. The amount of decline in price before a margin call is easy to calculate. Let P be the price that will result in a margin call. We will calculate this price for our original example where 100 shares of AT&T were purchased at $50 a share using $3,000 in cash and $2,000 in borrowed funds.

If the maintenance margin requirement is 25%, then

$$0.25 = \frac{100P - 2000}{100P}$$

and

$$P = 26\frac{2}{3}$$

Effect of Margin on Return

Margin is the purchase of securities utilizing leverage. As such, all gains and losses are accentuated. The amount of the accentuation depends on the percentage of the purchase the investor paid for in cash. Assume the share was purchased at $50. A $5 increase in price over six months would result in a six-month return for the security of

$$r_s = \frac{5}{50} = 10\%$$

Now assume the share was purchased with 50% margin and that the annual interest rate on the borrowing was 6% or 3% semiannually. With 50% margin the investor would put up $25 in cash, and the interest paid over the six months would be 0.03 (25) = $0.75. A $5 increase in share price results in a return on the cash investment (r_c) of

$$r_c = \frac{5 - 0.75}{25} = 17\%$$

Of course, this leverage works both ways. A $5 decrease in share value (share price declined 10%) would result in a percentage loss to the investor utilizing margin of

$$r_c = \frac{-5 - 0.75}{25} = -23\%$$

The minimum amount the investor puts up in cash is the margin times the price. In this case the return on the cash invested is

$$r_C = \frac{\text{Change in price} - \text{Interest}}{\text{Price} \times \text{Margin}} = \frac{1}{\text{Margin}} \frac{\text{Change in price} - \text{Interest}}{\text{Price}}$$
$$= \frac{1}{\text{Margin}} \left(r_S - \frac{\text{Interest}}{\text{Price}} \right)$$

There are also margin rules for short sales. The short seller receives the proceeds of the sale less the commission in the form of cash in the account. However, the short seller has to put up cash to protect against an increase in the stock price. Like long purchases, there are both margin requirements at the time of the trade (initial margin) and margin requirements to be met at all times (maintenance margin). The margin for short sales is calculated somewhat differently than the margin for purchases. This is the subject of the next section.

Margin Requirements for Short Sales

Margin for short sales is calculated as a percentage of the market value of the short. For short sales, margin is defined as

$$\text{Margin} = \frac{\text{Value of the assets} - \text{Market value of securities sold short}}{\text{Market value of securities sold short}}$$

For example, assume an investor just opened an account and short sold $10,000 worth of shares and the initial margin requirement was 50%. The investor would have to put up $5,000 in cash. The account would then look like this:

ASSETS		LIABILITIES	
Cash from the short sale	$10,000	Market value of securities sold short	$10,000
Cash from investor	$ 5,000	Net worth	$ 5,000
	$15,000		$15,000

and the account would meet the initial margin requirement because

$$\text{Margin} = \frac{15,000 - 10,000}{10,000} = 50\%$$

The amount of money that must be added to the account for additional short sales depends on what happened to share price subsequent to the short sales. If the stock price falls, then an account will be above the initial margin requirement, and additional shares can be sold short without putting up as much additional money, or perhaps no money at all. If the margin is below the initial margin, then the investor must bring the account up to the initial margin for additional short selling; therefore additional short sales involve more than a normal cash contribution. Short sales, like long purchases, require a minimum margin to be exceeded at all times; this is called a maintenance margin. Many accounts have both short sales and long purchases. These accounts would need to meet margin requirements for both types of trades.

MARKETS

In this section we discuss the markets in which trades take place. The section is divided into two parts. In the first part we discuss the general characteristics of markets. In the second part we discuss some of the principal U.S. markets.

Characteristics of Markets

There are a number of ways to classify markets. First, markets can be classified as primary or secondary. Primary markets are security markets where new issues of securities are initially sold. The Federal Reserve auctions off on a weekly basis new government bills and on a less frequent basis government bonds. This auction market is considered a primary market. A secondary market is a market where securities are resold. The New York Stock Exchange (NYSE) is a secondary market.

A second way to classify markets is as call or continuous markets. In a call markets, trading takes place at specified time intervals. One structure for a call market has prices announced verbally. In a verbal market, prices are announced, and the participants indicate the amount they are willing to sell or purchase at that price. This price is changed until a price is determined that most closely matches intended sales with intended purchases, at which time transactions are executed at that price.

A second structure for a call market uses a computer. Prices at which investors wish to buy or sell are entered into the computer, and a preliminary price is displayed. Investors can change their orders or enter new orders until a specified execution time when the price that best matches buys and sells is determined. If there is no price that completely matches buys and sells, an allocation method is needed. One method is first come, first served, which fills the oldest orders on the side with the surplus first.

Some call markets have a provision that limits the movement from the prior price. This is to prevent a temporary order imbalance from dramatically moving the price. Market orders are allowed in most call markets, and all market orders are filled at the clearing price. There is a greater price uncertainty for market orders in a call market than there is in a continuous market. In particular, the price movement between calls is likely to be greater than the price change in a continuous market from the time an order is placed until it is executed. Also, the trade need not be executed if the market has price limits and the clearing price exceeds the price limits. The NYSE opens the market with a trade very much like those found in a call market, though it then becomes a continuous market. Stock markets in Austria and Belgium are call markets, and Germany and Israel have call markets at some point in the day.

Continuous markets are markets where trading takes place on a continuous basis. For example, a market order placed in a continuous market will be executed quickly at the best available price.

A third way to classify markets is to determine whether they are dealer or broker markets. In a broker market, a broker acts as an agent for an investor and buys or sells shares on the investor's behalf. In a broker market, shareholders are trading with other shareholders, albeit utilizing an agent. In a dealer market, the dealer purchases or sells shares for the investor utilizing the dealer's own inventory. In a dealer market, investors' trades are not made directly with other investors but rather with the dealer, who serves as an intermediary between buyers and sellers.

A fourth way to classify markets is to determine whether the trading is executed by humans or done electronically. Execution on the NYSE involves people. Executions are done electronically on the Paris, Australia, and Toronto stock exchanges and for some stocks on the Tokyo stock exchange. One advantage of an electronic market is that the power of the computer allows complex conditional trades to be handled. For example, electronic trading would allow an order to be executed conditional on the value of a market index.

However we classify a market, there are a number of characteristics that are desirable for it to have. First, investors buy and sell assets based on information. Useful market information includes past prices, volume, current bids and offers, and the amount of short sales outstanding. Thus it is desirable that this market information be promptly and accurately

available to investors. Second, markets differ according to trading costs. The lower the costs of trading shares in the market, ceteris paribus, the better the market. Third, the markets should be liquid. *Liquidity* refers to the ability to transact a large number of shares at prices that don't vary substantially from past prices unless new information enters the market. Liquidity is often subdivided into *continuity* and *depth*. Price continuity means that an investor can expect to transact some shares at prices close to those at which the security recently traded absent any new information in the marketplace. A deep market is one that has a large number of buyers and sellers willing to trade at close to the current transaction price, so that a large number of shares can be transacted without a substantial change in price. Fourth, markets differ in the speed with which new information is incorporated into share prices. Investors would hope that the share price reflected all available information about the share. This is referred to as informational efficiency and is discussed in some detail in Chapter 17.

Major Markets

In this section we discuss some of the important markets in the United States, including both primary and secondary markets. First we discuss markets that are principally secondary markets.

Stock Markets Most stock trading in the United States takes place on exchanges. The most familiar are the NYSE (formally "NYSE Equities") and NASDAQ, but there are a number of others. An exchange provides two services: listing and trading. A listing is primarily a kind of sponsorship. The fact that IBM is listed on the NYSE implies that it meets certain standards of size, financial reporting, and governance. For this certification, IBM pays listing fees to the NYSE. Most U.S. firms list on one exchange: the NYSE, NASDAQ, NYSE MKT (formerly the American Stock Exchange), or NYSE ARCA. The listing exchange does not have a monopoly over trading. IBM can, and does, trade on many exchanges.

The trading services an exchange provides generally take the form of a computer platform governed by rules and procedures. The most widely used trading mechanism in equity markets is the *limit order book*, sometimes simply called a book or *order-driven market*. In a book market, customer limit orders that cannot be immediately executed (because the buy limit price is too low or the sell limit price is too high) are collected in a book. In the book, orders are ranked by price and time. That is, on the bid (buy) side of the book, a high-priced limit order has priority over a lower-priced order. On the offer (ask) side of the book, a low-priced limit order has priority. If two orders have the same price, the one that arrived at the exchange earlier has priority over the later arrival.

The highest bid and lowest offer prices in an exchange's book constitute the exchange's *best bid and offer* (BBO). A trade (execution) occurs when an incoming order is priced to meet either the best bid or the best offer. For example, suppose that stock XYZ is $20 bid and offered at $21; a buyer can trade immediately by bidding $21, and a seller can trade immediately by offering $20. The order on the book is called the *resting* or *passive order*. The incoming order is the "aggressor." The execution price is determined by the resting order. If the best bid (in the book) is $20, an incoming order to sell limit $3 will result in an execution at $20. An incoming order to sell limit $20.50 is not priced to hit the bid and so could not be immediately executed. It would be added to the book on the sell (offer) side. In this case, the added order improves on the prices of previous orders. The exchange's new best offer is $21.50.

Quantities matter, of course. Suppose that $20 is bid for 200 shares and that the next bid in the book is a $19.90 bid for 400 shares. An incoming order "sell 300 shares, limit $19" would be executed in two trades: 200 shares would be sold at $20 and 100 shares at $19.90. An incoming order that executes at multiple prices is said to "walk through the book." A market order to buy is treated as if it carries an infinite limit price. A buyer willing to pay "the market" has (in principle at least) no limit. A market order to sell is implicitly priced at zero. Most of the time, market orders will execute at prices close to other recent trades, but in turbulent markets the outcomes are quite volatile. During the May 6, 2010, "flash crash," Accenture (symbol ACN) traded at $41.52 per share at 2:30 PM, and at $0.01 per share at 2:47 PM. The penny-per-share trade occurred because a market order arrived when the best bid was $0.01. As it happened, ACN closed for the day above $40, and the $0.01 trade was broken (voided, by the exchange). The event nevertheless illustrates the dangers of unpriced orders, and many exchanges simply refuse to accept them.

Let us return to XYZ's, $20 bid offered at $21, and suppose that we want to buy. How should we price our order? "Buy limit $21" will usually give us an immediate execution. But we might do better on price if we are willing to wait. "Buy limit $20.90" establishes us as the new best bid, at a price much more attractive to potential sellers. A potential seller might view the improved bid as just good enough to meet and sell to us at our limit price. "Buy limit $20.01" also improves on the bid but is unlikely to encourage a prompt response. There is, of course, a danger in waiting. If XYZ makes a surprisingly positive earnings announcement, sellers are not likely to materialize at our price. It is more probable that sellers on the offer side of the book will withdraw their orders and resubmit them at higher prices. Other bidders (our competitors) will step up and surpass our bid. Our $20.90 bid is now priced "away from the market." We can chase the market, regretting our lost opportunity to buy at $21, or continue to wait in the hopes of subsequent decline.

The trade-off between price and execution certainty is fundamental to trading strategies and is difficult to assess. It is only one of the decisions confronting a trader. Another problem concerns the rate of trade. An institutional trader might be trying to buy or sell 100,000 shares of a stock that normally trades 10,000 shares per day. Sent to the market as a single, aggressively priced order, 100,000 shares will run through the book executing (at least in part) at very inferior prices. The price impact of the 100,000 share *parent order* can be minimized by working the order over time, feeding it to the market over several days as a series of smaller *child orders*. The calculations and analysis necessary to achieve optimal order splitting and pricing are sufficiently complex that the resulting strategies are implemented with automated processes and are described as *algorithmic*.

It was earlier noted that a stock can trade in many different exchanges. There is a limit order book for IBM at the NYSE (the firm's listing exchange). But there is also an IBM book at the BATS Exchange, at the DirectEdge Exchange, and in many other market centers. The exchanges will not necessarily all be showing the same best bid and offer, and there are other differences as well. There are differences in trading protocols: some exchanges allow users to hide their orders, there are small differences in fees to use the exchange (apart from the bid or offer nominally paid or received), and so on.

From a regulatory viewpoint, the diverse U.S. exchanges are considered to compose a "National Market System." The Securities and Exchange Commission's "Regulation National Market System" guides the permissible interactions across exchanges. All trades are reported to a consolidated system and are quickly made public. The best bid and offer of each exchange are similarly consolidated and disseminated. Exchanges are not supposed to *trade through* each other's visible quotes. That is, if the NYSE posts a best bid of $20 for XYZ, other exchanges cannot execute trades at prices below $20. In a non-NYSE

trade at $19, for example, the seller could have received a better price by routing the order to the NYSE. Moreover, the NYSE bidder at $20 is deprived of an execution.

Regulation NMS certainly simplifies the decisions faced by traders, but it does not solve all problems. There is no time priority, for example, across exchanges. A limit order placed on exchange A at 2:00 PM might execute before an identically priced order placed on exchange B at 10:00 AM. There are, furthermore, indeterminacies related to timing. When the NYSE is bidding $20, a trade elsewhere at $19 is normally considered a trade-through. But what if the NYSE's bid was posted one millisecond (one one-thousandth of a second) prior to the trade? Would the trader have been aware of the NYSE's bid?

The NYSE's data center is located in Mahwah, New Jersey. A trader in Chicago is about 1,100 kilometers away. At the speed of light, the round-trip delay is about seven milliseconds. Although this is trivial relative to human reaction times, a computerized algorithm operating in Chicago is at a significant disadvantage. The trader might not want to move to New Jersey, but she will certainly consider locating her computers there. The practice of placing the trading algorithms on computers in the same room as the Exchange's computers is commonly known as *collocation*. It is one technique for minimizing latency (delay).

High-frequency trading is generally characterized by automation, collocation, and various other practices that aggressively accelerate the reaction time to market developments. It is a controversial practice. Its practitioners claim that the technology helps tie markets together and lowers trading costs for other users. Detractors claim that high-frequency traders profit at the expense of other users and that the technology aggravates market instability.

Bond Markets Almost all bond trading in the secondary market occurs in the over-the-counter (OTC) market. While there is limited listing of bonds on the NYSE and AMEX, almost all the volume is OTC. The trading volume in government bonds is very large, and they are highly liquid. A number of government bond security dealers are willing to trade on a continuous basis, and most trading occurs with or between these dealers. An institution interested in trading in government bonds would call a number of government bond dealers and get quotes. The institution would then take one of the quotes or attempt to negotiate a more attractive price. An individual purchasing through a broker would be offered whatever the current bid and ask are from that broker for retail accounts. Treasury dealers trade with each other in a different manner using intermediaries called government brokers. Five government brokers handle the majority of the trading volume. Treasury dealers give firm bids and offers to the government broker who displays the most attractive bid and offer on a monitor at each dealer. Dealers can execute the trade electronically, and the size and price of the trade are immediately available to all dealers. The government bond brokers deal with all the paperwork and maintain the confidentiality of the traders in return for a small fee. Thus, although there is a quote system for dealers, none exists for individuals or nondealer institutions. In addition, there is no record of transaction prices available to the general public.

The secondary markets for corporate bonds, or Ginnie Maes, are fairly illiquid. Only recent issues or some large issues have an active secondary market. An individual wishing to purchase a bond with certain characteristics will likely be offered a choice of bonds with these characteristics that are contained in the brokerage firm’s inventory. Given the illiquidity, the characteristics will have to be stated in fairly general terms (e.g., AAA corporates with about 10 years’ maturity). If the inquiry involves a sufficiently large order, the firm might survey other firms to determine other potential options. Given the illiquidity of the market, the bid–ask spread will be much higher than in the government bond market.

Dealers have three sources of potential profit: (1) they can make money on the bid–ask spread, (2) they can make or lose money on the change in the value of the inventory,[6] and (3) they can make or lose money on the difference between the interest earned on the inventory and the interest paid to finance it.

Primary Markets Primary markets are markets that involve new issues of securities and hence, unlike secondary markets, provide a direct flow of cash to the issuing entity. In this section we discuss some of the principal primary markets.

Government Bonds Treasury securities are issued by auction on a regular basis, where the frequency of issuance depends on the maturity of the security. For example, 91-day and 182-day Treasury bills are offered every Monday; 7-, 10-, and 30-year Treasury bonds are issued quarterly. Two types of orders can be placed: noncompetitive (market orders) and competitive (limit orders). Noncompetitive bids can be placed up to $1 million face value. Noncompetitive bids are filled at a price equal to the average price paid by all competitive bidders.

Competitive bids can be placed by banks or brokerage firms that are designated by the Federal Reserve. These institutions place bids for a particular quantity and at a particular price (limit orders) for themselves or their customers. The auction works as follows. First, the Federal Reserve deducts the aggregate value of all noncompetitive issues from the aggregate amount to be sold. It then ranks the competitive bids from highest to lowest, filling the bids until the amount it wishes to issue is sold. For the marginal bids (the lowest accepted) the volume is allocated among bidders proportional to the amount requested by each. Thus competitive bidders can receive the amount they bid, a fractional amount, or none. Noncompetitive bidders have price uncertainty; competitive bidders face volume uncertainty.

Corporate Issues Corporate bonds and common stocks are usually sold using the services of an investment banker. The corporation normally has an ongoing relationship with an investment banker; when it has a need for funds, the corporation negotiates the instruments and price with the investment banker. New issues are divided into two types: (1) seasoned new issues, which are issues of companies that already have publicly traded securities, such as new issues of Ford or IBM; and (2) issues of companies without publicly traded securities, referred to as initial public offerings (IPOs). These issues are usually issues of small companies just starting out. However, they can be issues of companies that are recapitalizing, such as companies that had publicly traded securities, were bought out by the management and held privately, and then became public again. These companies can be quite large (e.g., Nabisco).

The investment banking firm either can purchase the shares directly from the firm at an agreed-upon price and then resell them to the general public (called *firm commitment*) or can simply help the firm in selling to the general public (called *best efforts*). Underwriters have a conflict of interest. As an adviser to the issuing firm, they have an obligation to obtain the best price possible. However, the lower the price, the easier it is to market the securities to the public. The empirical evidence indicates that IPOs earn abnormally large returns on the day of issuance but underperform similar-risk securities in subsequent months.

Clearing Procedures Most transactions require that settlement be made in five business days. A brokerage firm will engage in trades involving customers from a number of other brokerage firms in the same security. Some will be sales, and some will be purchases.

[6]Many dealers will try to limit the susceptibility to price changes on inventory by taking an appropriate position in the futures market (for a discussion of how this is done, see Chapter 24).

It would be very costly to have to settle each and every trade rather than the net of the purchases and sales. To facilitate settlement, clearing corporations have been established. At the end of the day, all records of trades are sent to the clearing corporation, which then notifies the firm of the net amount of securities to be delivered and the net amount of money to be received or to pay.

Clearing corporations play an especially important role in the options and futures markets. Not only are all trades cleared through the clearing corporation, the clearing corporation guarantees all trades. With options and futures (unlike common stock), the profit or loss comes from the individual on the other side of the trade. With stocks and bonds, the profit depends on the creditworthiness of a corporation and its earnings, and one can analyze the creditworthiness of a publicly held corporation. It would be much more difficult to determine the creditworthiness of the individual taking the opposite position in an option or a future. In these markets the clearing corporation stands behind each trade, and thus the trade can be thought of as a trade with the corporation. The clearing corporation keeps a list of all buyers and sellers of each security, but there is no matching of trades. If a trader fails to meet his or her obligation, margin is taken, and the firm that executed the trade makes up any difference. If the firm that made the trade fails, there is a further system of backup involving the failing firm's margin, the margin of other firms, and the clearing corporation's own assets. In short, the risk of an investor in the options and futures markets not having his or her contract honored is the risk of the clearing corporation collapsing, and such a collapse would have to involve massive failure of much of the financial system.

The clearing corporation serves another role in the options and futures market. Because trades can be thought of as taking place with the clearing corporation, when the investor exercises an option or delivers on a futures contract, the clearing corporation has to decide who is on the other side. This is decided by using well-specified rules, which in some cases are random selection.

TRADE TYPES AND COSTS

In this section we discuss motivations for trading and what factors influence the costs of a trade.

Types of Trades

There are generally considered to be two reasons for investors to trade. The first reason investors trade securities is that traders believe that the price is incorrect, and they buy or sell based on a perceived mispricing. These traders are referred to as information traders. The second reason for buying or selling securities is because of a surplus of or need for money. An investor needing the down payment for a house or the money to purchase a car or boat might liquidate part of his or her portfolio to obtain the necessary funds. Similarly, an investor receiving an inflow of funds might purchase shares because stocks in general are a good investment rather than because of information indicating that the particular stock being purchased is mispriced. These investors are referred to as liquidity traders.

Specialists and dealers have different profit possibilities trading with information-based or liquidity-based traders. For liquidity-based traders, it is reasonable to assume that the side of the trade they are on (buy or sell) is unrelated to the future course of price movements. Thus the bid–ask spread should provide profit to the dealer, because subsequent price movements will not systematically affect the value of any inventory held as a result of the trade. This is true whether we are discussing a specialist buying or selling for his or her own account or a dealer in the government bond market. An information-based trade

is different. If the person initiating the trade has superior information, then one can anticipate that the short-run price movements on average will be unfavorable to the specialist or dealer. And although the specialist will gain on the bid–ask spread, he or she will lose on any inventory obtained or lost from the trade because subsequent price movements will likely be unfavorable on average. The specialist or dealer will expect to gain money from liquidity traders and information traders who do not have superior information but to lose to information traders with superior information. The greater the proportion and the higher the quality of the superior information for information-based traders, the more the specialist or dealer will have to make on the bid–ask spread and the higher the bid–ask spread must be. Thus liquidity investors will do better if specialists and dealers are informed. Furthermore, a liquidity-based trader who can credibly convey this fact to the dealers should be able to obtain a better price. For example, an index fund that stays fully invested can often obtain better prices in purchases and sales because its trades are not information motivated, which is a credible message to convey to dealers.

Trading Costs

One of the important elements in markets is the cost of trading. The size of the trading costs affects how large the perceived mispricing must be before an investor can profitably swap one share for another. Substantial trading costs mean that investors will hold nonoptional portfolios because the transaction costs of adjusting them are too high.

There are three major sources of trading costs. First are the direct costs: commission to the brokers plus a tax on the trade. The second cost is the bid–ask spread. An investor buying and then subsequently selling the stock will purchase at the ask and sell at the bid. The difference in the bid and ask is a cost to the investor buying and then selling the stock (called *roundtrip*). Third is the potential price impact of a large sale or purchase. Small purchases and sales can be executed at the bid and ask, but large purchases or sales may cause an adverse change in the bid and ask.

There is another factor that affects the cost for liquidity traders. Because liquidity traders do not engage in a determination of equilibrium price, it is important that they feel confident that market prices are close to equilibrium prices. Quoted prices can differ from equilibrium prices because information takes a long time to be incorporated into a share price, or because trading costs are sufficiently high that trade prices can differ substantially from equilibrium prices without information-based traders entering the market. Differences of trade prices from equilibrium prices can help or hurt the investor's return. However, these differences increase the variability of return the investor will receive and thus are a cost in the sense that they increase the investor's risk.

These costs vary with the type of security purchased or sold, the exchange used (if any), and the size of the purchase or sale. Commissions as a percentage of the total value of the sale generally increase as the price of the share declines. The commission also varies widely from broker to broker. Full-service brokerage firms offering advice as well as transaction services generally charge substantially more than the discount brokers, who primarily offer order execution. The bid–ask spread also varies across securities. The less liquid the security, the greater the bid–ask spread is likely to be. This is especially true in the bond market, where very illiquid bonds are likely to have very large bid–ask spreads.

The size of the trade works both ways. The larger the trade, the more likely the trade is to have an impact on price. Large traders, however, usually institutions, are in a better position to negotiate a more attractive price. This is especially true in dealer markets, such as bond markets, where large investors have the ability to negotiate with a number of dealers and where small investors usually do not have access to negotiation with multiple dealers.

CONCLUSION

In this chapter we have described the markets in which securities are traded. This chapter and Chapter 2 have supplied the reader with the background necessary for the discussion of investment analysis. We start with a more detailed analysis of risk and portfolio management. This will allow us to return to an examination of many of the securities discussed to this point to see how they are priced and how they fit into a portfolio.

Part 2

PORTFOLIO ANALYSIS

Section 1

Mean Variance Portfolio Theory

4

The Characteristics of the Opportunity Set under Risk

In Chapter 1 we introduced the elements of a decision problem under certainty. The same elements are present when we recognize the existence of risk; however, their formulation becomes more complex. In the next two chapters we explore the nature of the opportunity set under risk. Before we begin the analysis we present a brief summary or road map of where we are going. The existence of risk means that the investor can no longer associate a single number or payoff with investment in any asset. The payoff must be described by a set of outcomes and each of their associated probabilities of occurrence, called a frequency function or return distribution. In this chapter we start by examining the two most frequently employed attributes of such a distribution: a measure of central tendency, called the expected return, and a measure of risk or dispersion around the mean, called the standard deviation. Investors should not and, in fact, do not hold single assets; they hold groups or portfolios of assets. Thus a large part of this chapter is concerned with how one can compute the expected return and risk of a portfolio of assets given the attributes of the individual assets. One important aspect of this analysis is that the risk on a portfolio is more complex than a simple average of the risk on individual assets. It depends on whether the returns on individual assets tend to move together or whether some assets give good returns when others give bad returns. As we show in great detail, there is a risk reduction from holding a portfolio of assets if assets do not move in perfect unison.

We continue this discussion in Chapter 5. Initially, we examine portfolios of only two assets. We present a detailed geometric and algebraic analysis of the characteristics of portfolios of two assets under different estimates of how they covary together (how related their returns are to each other). We then extend this analysis to the case of multiple assets. Finally, we arrive at the opportunity set facing the investor in a world with risk. Let us begin by characterizing the nature of the opportunity set open to the investor.

In the certainty case, the investor's decision problem can be characterized by a certain outcome. In the problem analyzed in Chapter 1, the 5% return on lending (or the 5% cost of borrowing) was known with certainty. Under risk, the outcome of any action is not known with certainty, and outcomes are usually represented by a frequency function. A frequency function is a listing of all possible outcomes along with the probability of the occurrence of each. Table 4.1 shows such a function. This investment has three possible

returns. If event 1 occurs, the investor receives a return of 12%; if event 2 occurs, 9% is received; and if event 3 occurs, 6% is received. In our examples each of these events is assumed to be equally likely. Table 4.1 shows us everything there is to know about the return possibilities.

Table 4.1 Data on Three Hypothetical Events

Return	Probability	Event
12	$\frac{1}{3}$	1
9	$\frac{1}{3}$	2
6	$\frac{1}{3}$	3

Usually we do not delineate all of the possibilities, as we have in Table 4.1. The possibilities for real assets are sufficiently numerous that developing a table like Table 4.1 for each asset is too complex a task. Furthermore, even if the investor decided to develop such tables, the inaccuracies introduced would be so large that he or she would probably be better off just trying to represent the possible outcomes in terms of some summary measures. In general, it takes at least two measures to capture the relevant information about a frequency function: one to measure the average value and one to measure dispersion around the average value.

DETERMINING THE AVERAGE OUTCOME

The concept of an average is standard in our culture. Pick up the newspaper and you will often see figures on average income, batting averages, or average crime rates. The concept of an average is intuitive. If someone earns $11,000 one year and $9,000 in a second, we say his average income in the two years is $10,000. If three children in a family are age 15, 10, and 5, then we say the average age is 10. In Table 4.1 the average return was 9%. Statisticians usually use the term *expected value* to refer to what is commonly called an average. In this book we use both terms.

An expected value or average is easy to compute. If all outcomes are equally likely, then to determine the average, one adds up the outcomes and divides by the number of outcomes. Thus, for Table 4.1, the average is $(12 + 9 + 6)/3 = 9$. A second way to determine an average is to multiply each outcome by the probability that it will occur. When the outcomes are not equally likely, this facilitates the calculation. Applying this procedure to Table 4.1 yields $\frac{1}{3}(12) + \frac{1}{3}(9) + \frac{1}{3}(6) = 9$.

It is useful to express this intuitive calculation in terms of a formula. The symbol Σ should be read "sum." Underneath the symbol we put the first value in the sum and what is varying. On the top of the symbol we put the final value in the sum. We use the symbol R_{ij} to denote the jth possible outcome for the return on security i. Thus

$$\frac{\sum_{j=1}^{3} R_{ij}}{3} = \frac{R_{i1} + R_{i2} + R_{i3}}{3} = \frac{12 + 9 + 6}{3}$$

Using the summation notation just introduced and a bar over a variable to indicate expected return, we have for the expected value of the M equally likely returns for asset *i*:

$$\bar{R}_i = \sum_{j=1}^{M} \frac{R_{ij}}{M}$$

If the outcomes are not equally likely and if P_{ij} is the probability of the *j*th return on the *i*th asset, then expected return is[1]

$$\bar{R}_i = \sum_{j=1}^{M} P_{ij} R_{ij}$$

We have up to this point used a bar over a symbol to indicate expected value. This is the procedure we adopt throughout most of this book. However, occasionally, this notation proves awkward. An alternative method of indicating expected value is to put the symbol E in front of the expression for which we wish to determine the expected value. Thus $E(R_i)$ should be read as the expected value of R_{ij}, just as $\bar{R}_i$ is the expected value of R_{ij}.

Certain properties of expected value are extremely useful:

1. The expected value of the sum of two returns is equal to the sum of the expected value of each return, that is,

$$E\left(R_{1j} + R_{2j}\right) = \bar{R}_1 + \bar{R}_2$$

2. The expected value of a constant C times a return is the constant times the expected return, that is,

$$E\left[C\left(R_{1j}\right)\right] = C\bar{R}_1$$

These principles are illustrated in Table 4.2. For any event, the return on asset 3 is the sum of the return on assets 1 and 2. Thus the expected value of the return on asset 3 is the sum of the expected value of the return on assets 1 and 2. Likewise, for any event, the return on asset 3 is 3 times the return on asset 1. Consequently, its expected value is 3 times as large as the expected value of asset 1.

These two properties of expected values will be used repeatedly and are worth remembering.

Table 4.2 Return on Various Assets

Event	Probability	Asset 1	Asset 2	Asset 3
A	$\frac{1}{3}$	14	28	42
B	$\frac{1}{3}$	10	20	30
C	$\frac{1}{3}$	6	12	18
	Expected return	10	20	30

A MEASURE OF DISPERSION

Not only is it necessary to have a measure of the average return but it is also useful to have some measure of how much the outcomes differ from the average. The need for this second

[1]This latter formula includes the formula for equally likely observations as a special case. If we have M observations each equally likely, then the odds of any one occurring are $1/M$. Replacing the P_{ij} in the second formula with $1/M$ yields the first formula.

characteristic can be illustrated by the old story of the mathematician who believed an average by itself was an adequate description of a process and drowned in a stream with an average depth of 2 inches.

Intuitively, a sensible way to measure how much the outcomes differ from the average is simply to examine this difference directly; that is, examine $R_{ij} - \bar{R}_i$. Having determined this for each outcome, one could obtain an overall measure by taking the average of this difference. Although this is intuitively sensible, there is a problem. Some of the differences will be positive and some negative, and these will tend to cancel out. The result of the canceling could be such that the average difference for a highly variable return need be no larger than the average difference for an asset with a highly stable return. In fact, it can be shown that the average value of this difference must always be precisely zero. The reader is encouraged to verify this with the example in Table 4.2. Thus the sum of the differences from the mean tells us nothing about dispersion.

Two solutions to this problem suggest themselves. First, we could take absolute values of the difference between an outcome and its mean by ignoring minus signs when determining the average difference. Second, because the square of any number is positive, we could square all differences before determining the average. For ease of computation, when portfolios are considered, the latter procedure is generally followed. In addition, as we will see when we discuss utility functions, the average squared deviations have some convenient properties.[2] The average squared deviation is called the *variance*; the square root of the variance is called the *standard deviation*. In Table 4.3 we present the possible returns from several hypothetical assets as well as the variance of the return on each asset. The alternative returns on any asset are assumed equally likely. Examining asset 1, we find the deviations of its returns from its average return are $(15 - 9)$, $(9 - 9)$, and $(3 - 9)$. The squared deviations are 36, 0, and 36, and the average squared deviation or variance is $(36 + 0 + 36)/3 = 24$.

To be precise, the formula for the variance of the return on the ith asset (which we symbolize as σ_i^2) when each return is equally likely is

$$\sigma_i^2 = \sum_{j=1}^{M} \frac{\left(R_{ij} - \bar{R}_i\right)^2}{M}$$

Table 4.3 Returns on Various Investments[a]

Market Condition	Return[a] Asset 1	Asset 2	Asset 3	Asset 5	Rainfall	Return[a] Asset 4
Good	15	16	1	16	Plentiful	16
Average	9	10	10	10	Average	10
Poor	3	4	19	4	Poor	4
Mean return	9	10	10	10		10
Variance	24	24	54	24		24
Standard deviation	4.9	4.9	7.35	4.90		4.9

[a]The alternative returns on each asset are assumed equally likely, and thus each has a probability of $\frac{1}{3}$.

[2]Many utility functions can be expressed either exactly or approximately in terms of the mean and variance. Furthermore, regardless of the investor's utility function, if returns are normally distributed, the mean and variance contain all relevant informaton about the distribution. An elaboration of these points is contained in later chapters.

If the observations are not equally likely, then, as before, we multiply by the probability with which they occur. The formula for the variance of the return on the *i*th asset becomes

$$\sigma_i^2 = \sum_{j=1}^{M} \left[P_{ij} \left(R_{ij} - \overline{R}_i \right)^2 \right]$$

Occasionally, we will find it convenient to employ an alternative measure of dispersion called standard deviation. The standard deviation is just the square root of the variance and is designated by σ_i.

In the examples discussed in this chapter we are assuming that the investor is estimating the possible outcomes and the associated probabilities. Often initial estimates of the variance are obtained from historical observations of the asset's return. In this case, many authors and programs used in calculators multiply the variance formula given earlier by $M/(M - 1)$. This produces an estimate of the variance that is unbiased but has the disadvantage of being inefficient (i.e., it produces a poorer estimate of the true variance). We leave it to readers to choose which they prefer. In our examples in this book, we will not make this correction.[3]

The variance tells us that asset 3 varies considerably more from its average than asset 2. This is what we intuitively see by examining the returns shown in Table 4.3. The expected value and variance or standard deviation are the usual summary statistics utilized in describing a frequency distribution.

There are other measures of dispersion that could be used. We have already mentioned one, the average absolute deviation. Other measures have been suggested. One such measure considers only deviations below the mean. The argument is that returns above the average return are desirable. The only returns that disturb an investor are those below average. A measure of this is the average (overall observations) of the squared deviations below the mean. For example, in Table 4.3, for asset 1, the only return below the mean is 3. Because 3 is 6 below the mean, the square of the difference is 36. The other two returns are not below the mean, so they have 0 deviation below the mean. The average of (0) + (0) + (36) is 12. This measure is called the *semivariance*.

Semivariance measures downside risk relative to a benchmark given by expected return. It is just one of a number of possible measures of downside risk. More generally, we can consider returns relative to other benchmarks, including a risk-free return or zero return. These generalized measures are, in aggregate, referred to as *lower partial moments*. Yet another measure of downside risk is the so-called *value at risk measure*, which is widely used by banks to measure their exposure to adverse events and to measure the least expected loss (relative to zero, or relative to wealth) that will be expected with a certain probability. For example, if 5% of the outcomes are below −30%, and if the decision maker is concerned about how poor the outcomes are 5% of the time, then −30% is the value at risk.

[3]As stated, sometimes the formula is divided by *M*, and sometimes it is divided by $M - 1$. The choice is a matter of taste. However, the reader may be curious why some choose one or the other. The technical reason authors choose one or the other is as follows.

Employing *M* as the denominator gave the best estimate of the true value or the so-called maximum likelihood estimate. Although it is the best estimate as *M* gets large, it does not converge to the true value (it is too small). Dividing by $M - 1$ produces a S_i^2 that converges to the true value as *M* gets large (technically unbiased) but is not the best estimate for a finite *M*. Some people consider one of these properties more important than the other, whereas some use one without consciously realizing why this might be preferred.

Intuitively, these alternative measures of downside risk are reasonable, and some portfolio theory has been developed using them. However, they are difficult to use when we move from single assets to portfolios. In cases where the distribution of returns is symmetrical, the ordering of portfolios in mean variance space will be the same as the ordering of portfolios in mean semivariance space or mean and any of the other measures of downside risk discussed earlier. For well-diversified equity portfolios, symmetrical distribution is a reasonable assumption, so variance is an appropriate measure of downside risk. Furthermore, because empirical evidence shows most assets existing in the market have returns that are reasonably symmetrical, semivariance is not needed. If returns on an asset are symmetrical, the semivariance is proportional to the variance. Thus, in most of the portfolio literature, the variance, or equivalently the standard deviation, is used as a measure of dispersion.

In most cases, instead of using the full frequency function such as that presented in Table 4.1, we use the summary statistics mean and variance or equivalent mean and standard deviation to characterize the distribution. Consider two assets. How might we decide which we prefer? First, intuitively, one would think that most investors would prefer the one with the higher expected return if standard deviation was held constant. Thus, in Table 4.3, most investors would prefer asset 2 to asset 1. Similarly, if expected return were held constant, investors would prefer the one with the lower variance. This is reasonable because the smaller the variance, the more certain an investor is that she will obtain the expected return, and the fewer poor outcomes she has to contend with.[4] Thus, in Table 4.3, the investor would prefer asset 2 to asset 3.

VARIANCE OF COMBINATIONS OF ASSETS

This simple analysis has taken us partway toward an understanding of the choice between risky assets. However, the options open to an investor are not simply to pick between assets 1, 2, 3, 4, or 5 in Table 4.3 but also to consider combinations of these five assets. For example, an investor could invest part of her money in each asset. While this opportunity vastly increases the number of options open to the investor and hence the complexity of the problem, it also provides the raison d'être of portfolio theory. The risk of a combination of assets is very different from a simple average of the risk of individual assets. Most dramatically, the variance of a combination of two assets may be less than the variance of either of the assets themselves. In Table 4.4, there is a combination of asset 2 and asset 3 that is less risky than asset 2.

Table 4.4 Dollars at Period 2 Given Alternative Investments

Condition of Market	Asset 2	Asset 3	Combination of Asset 2 (60%) and Asset 3 (40%)
Good	$1.16	$1.01	$1.10
Average	1.10	1.10	1.10
Poor	1.04	1.19	1.10

[4]We will not formally develop the criteria for making a choice from among risky opportunities until the next chapter. However, we feel we are not violating common sense by assuming at this time that investors prefer more to less and act as risk avoiders. More formal statements of the properties of investor choice will be taken up in the next chapter.

Let us examine this property. Assume an investor has \$1 to invest. If he selects asset 2 and the market is good, he will have at the end of the period \$1 + 0.16 = \$1.16. If the market's performance is average, he will have \$1.10, and if it is poor, \$1.04. These outcomes are summarized in Table 4.4, along with the corresponding values for the third asset. Consider an alternative. Suppose the investor invests \$0.60 in asset 2 and \$0.40 in asset 3. If the condition of the market is good, the investor will have \$0.696 at the end of the period from asset 2 and \$0.404 from asset 3, or \$1.10. If the market conditions are average, he will receive \$0.66 from asset 2, \$0.44 from asset 3, or a total of \$1.10. By now the reader might suspect that if the market condition is poor, the investor still receives \$1.10, and this is, of course, the case. If the market condition is poor, the investor receives \$0.624 from his investment in asset 2 and \$0.476 from his investment in asset 3, or \$1.10. These possibilities are summarized in Table 4.4.

This example dramatically illustrates how the risk of a portfolio of assets can differ from the risk of the individual assets. The deviations on the combination of the assets were zero because the assets had their highest and lowest returns under opposite market conditions. This result is perfectly general and not confined to this example. When two assets have their good and poor returns at opposite times, an investor can always find *some* combination of these assets that yields the same return under all market conditions. This example illustrates the importance of considering combinations of assets rather than just the assets themselves and shows how the distribution of outcomes on combinations of assets can be different than the distributions on the individual assets.

The returns on asset 2 and asset 4 have been developed to illustrate another possible situation. Asset 4 has three possible returns. Which return occurs depends on rainfall. Assuming that the amount of rainfall that occurs is independent of the condition of the market, then the returns on assets 2 and 4 are independent of one another. Therefore, if the rainfall is plentiful, we can have good-, average-, or poor-security markets. Plentiful rainfall does not change the likelihood of any particular market condition occurring. Consider an investor with \$1.00 who invests \$0.50 in each asset. If rain is plentiful, he receives \$0.58 from his investment in asset 4 and any one of three equally likely outcomes from his investment in asset 2: \$0.58 if the market is good, \$0.55 if it is average, and \$0.52 if the market is poor. This gives him a total of \$1.16, \$1.13, or \$1.10. Similarly, if the rainfall is average, the value of his investment in assets 2 and 4 is \$1.13, \$1.10, or \$1.07, and if rainfall is poor, \$1.10, \$1.07, or \$1.04. Because we have assumed that each possible level of rainfall is equally likely as each possible condition of the market, there are nine equally likely outcomes. Ordering them from highest to lowest, we have \$1.16, \$1.13, \$1.13, \$1.10, \$1.10, \$1.10, \$1.07, \$1.07, and \$1.04. Compare this to an investment in asset 2 by itself, the results of which are shown in Table 4.3. The mean is the same; however, the dispersion around the mean is less. This can be seen by direct examination and by noting that the probability of one of the extreme outcomes occurring (\$1.16 or \$1.04) has dropped from $\frac{1}{3}$ to $\frac{1}{9}$.

This example once again shows how the characteristics of the portfolio can be very different than the characteristics of the assets that compose the portfolio. The example illustrates a general principle. When the returns on assets are independent, such as the returns on assets 2 and 4, a portfolio of such assets can have less dispersion than either asset.

Consider still a third situation, one with a different outcome than the previous two. Consider an investment in assets 2 and 5. Assume the investor invests \$0.50 in asset 2 and \$0.50 in asset 5. The value of his investment at the end of the period is \$1.16, \$1.10, or \$1.04. These are the same values he would have obtained had he invested the entire

$1.00 in either asset 2 or asset 5 (see Table 4.3). Thus, in this situation, the characteristics of the portfolios were exactly the same as the characteristics of the individual assets, and holding a portfolio rather than the individual assets did not change the investor's risk.

We have analyzed three extreme situations. As extremes, they dramatically illustrated some general principles that carry over to less extreme situations. Our first example showed that when assets have their good and bad outcomes at different times (assets 2 and 3), investment in these assets can radically reduce the dispersion obtained by investing in one of the assets by itself. If the good outcomes of an asset are not always associated with the bad outcomes of a second asset, but the general tendency is in this direction, then the reduction in dispersion still occurs, but the dispersion will not drop all the way to zero, as it did in our example. However, it is still often true that appropriately selected combinations of the two assets will have less risk than the least risky of the two assets.

Our second example illustrated the situation where the conditions leading to various returns were different for the two assets. More formally, this is the area where returns are independent. Once again, dispersion was reduced, but not in as drastic a fashion. Note that investment in asset 2 alone can result in a return of $1.04 and that this result occurs one-third of the time. The same result can occur when we invest an equal amount in asset 2 and asset 4. However, a combination of assets 2 and 4 has nine possible outcomes, each equally likely, and $1.04 occurs only one-ninth of the time. With independent returns, extreme observations can still occur; they just occur less frequently. Just as the extreme values occur less frequently, outcomes closer to the mean become more likely, so that the frequency function has less dispersion.

Finally, our third example illustrated the situation where the assets being combined had their outcomes affected in the same way by the same events. In this case, the characteristics of the portfolio were identical to the characteristics of the individual assets. In less extreme cases this is no longer true. Insofar as the good and bad returns on assets tend to occur at the same time, but not always exactly at the same time, the dispersion on the portfolio of assets is somewhat reduced relative to the dispersion on the individual assets.

We have shown with some simple examples how the characteristics of the returns on portfolios of assets can differ from the characteristics of the returns on individual assets. These were artificial examples designed to dramatically illustrate the point. To reemphasize this point, it is worthwhile examining portfolios of some real securities over a historical period.

Three securities were selected: Microsoft, Dell, and General Electric (G.E.). The monthly returns, average return, and standard deviation from investing in each security are shown in Table 4.5. In addition, the return and risk of placing one-half of the available funds in each pair of securities are shown in the table. Finally, we have plotted the returns from each possible pair of securities in Figure 4.1. In this figure we have the return from each of three securities as well as the return from placing one-half of the available funds in each pair of securities. Both Figure 4.1 and Table 4.5 make it clear how diversification across real securities can have a tremendous payoff for the investor. For example, a portfolio composed of 50% Dell and 50% G.E. had a higher return than G.E. but a lower risk. Earlier we argued that an investor is better off working with summary characteristics rather than full frequency functions. We used two summary measures: average return and variance or standard deviation of return. We now examine analytically how the summary characteristics of a portfolio are related to those of individual assets.

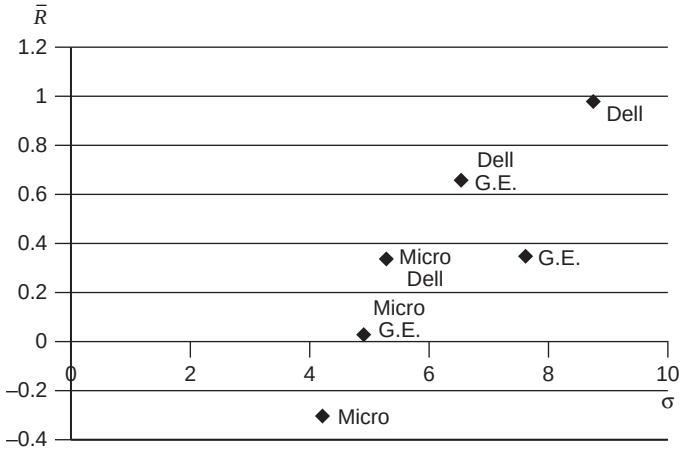


Figure 4.1 Securities and predetermined portfolios.

CHARACTERISTICS OF PORTFOLIOS IN GENERAL

The return on a portfolio of assets is simply a weighted average of the return on the individual assets. The weight applied to each return is the fraction of the portfolio invested in that asset. If R_{Pj} is the jth return on the portfolio and X_i is the fraction of the investor's funds invested in the ith asset, and N is the number of assets, then

$$R_{Pj} = \sum_{i=1}^{N} \left(X_i R_{ij} \right)$$

Table 4.5 Monthly Returns on Microsoft, Dell, and G.E. (in percent, 2011)

Month	Micro	Dell	G.E.	$\frac{1}{2}$ Micro + $\frac{1}{2}$ Dell	$\frac{1}{2}$ Micro + $\frac{1}{2}$ G.E.	$\frac{1}{2}$ Dell + $\frac{1}{2}$ G.E.
Jan.	−0.66%	−2.88%	10.11%	−1.77%	4.73%	3.62%
Feb.	−3.55%	20.29%	4.57%	8.37%	0.51%	12.43%
March	−4.48%	−8.34%	−4.16%	−6.41%	−4.32%	−6.25%
April	2.09%	6.62%	2.00%	4.35%	2.04%	4.31%
May	−2.89%	3.94%	−3.96%	0.52%	−3.43%	−0.01%
June	3.96%	3.67%	−3.21%	3.81%	0.38%	0.23%
July	5.38%	−2.58%	−5.04%	1.40%	0.17%	−3.81%
Aug.	−2.34%	−8.47%	−8.93%	−5.40%	−5.63%	−8.70%
Sept.	−6.43%	−4.88%	−5.76%	−5.65%	−6.10%	−5.32%
Oct.	6.99%	11.81%	9.79%	9.40%	8.39%	10.80%
Nov.	−3.19%	−0.32%	−4.79%	−1.75%	−3.99%	−2.55%
Dec.	1.49%	−7.17%	13.64%	−2.84%	7.56%	3.23%
Average	−0.30%	0.98%	0.35%	0.34%	0.03%	0.66%
Stdev	4.24%	8.76%	7.46%	5.30%	4.95%	6.55%

Correlation Coefficient
Microsoft and Dell = 0.24
Microsoft and G.E. = 0.39
Dell and G.E. = 0.30

The expected return is also a weighted average of the expected returns on the individual assets. Taking the expected value of the expression just given for the return on a portfolio yields

$$\bar{R}_P = E(R_P) = E\left(\sum_{i=1}^{N} X_i R_{ij}\right)$$

But we already know that the expected value of the sum of various returns is the sum of the expected values. Therefore, we have

$$\bar{R}_P = \sum_{i=1}^{N} E\left(X_i R_{ij}\right)$$

Finally, the expected value of a constant times a return is a constant times the expected return, or

$$\bar{R}_P = \sum_{i=1}^{N} \left(X_i \bar{R}_i\right)$$

This is a perfectly general formula, and we use it throughout the book. To illustrate its use, consider the investment in assets 2 and 3 discussed earlier in Table 4.3. We determined that no matter what occurred, the investor would receive \$1.10 on an investment of \$1.00. This is a return of 0.10/1.00 = 10%.

Let us apply the formula for expected return. In the example discussed earlier, \$0.60 was invested in asset 2 and \$0.40 in asset 4; therefore, the fraction invested in asset 4 is 0.40/1.00. Furthermore, the expected return on asset 2 and asset 4 is 10%. Applying the formula for expected return on a portfolio yields

$$\bar{R}_P = \left(\frac{0.60}{1.00}\right)(0.10) + \left(\frac{0.40}{1.00}\right)(0.10) = 0.10$$

The second summary characteristic was the variance. The variance on a portfolio is a little more difficult to determine than the expected return. We start out with a two-asset example. The variance of a portfolio P, designated by σ_P^2, is simply the expected value of the squared deviations of the return on the portfolio from the mean return on the portfolio, or $\sigma_P^2 = E(R_P - \bar{R}_P)^2$. Substituting in this expression the formulas for return on the portfolio and mean return yields in the two-security case

$$\begin{aligned}\sigma_P^2 = E\left(R_P - \bar{R}_P\right)^2 &= E\left[X_1 R_{1j} + X_2 R_{2j} - \left(X_1 \bar{R}_1 + X_2 \bar{R}_2\right)\right]^2 \\ &= E\left[X_1\left(R_{1j} - \bar{R}_1\right) + X_2\left(R_{2j} - \bar{R}_2\right)\right]^2\end{aligned}$$

where $\bar{R}_i$ stands for the expected value of security i with respect to all possible outcomes. Recall that

$$(X+Y)^2 = X^2 + XY + XY + Y^2 = X^2 + 2XY + Y^2$$

Applying this to the previous expression, we have

$$\sigma_P^2 = E\left[X_1^2\left(R_{1j} - \bar{R}_1\right)^2 + 2X_1X_2\left(R_{1j} - \bar{R}_1\right)\left(R_{2j} - \bar{R}_2\right) + X_2^2\left(R_{2j} - \bar{R}_2\right)^2\right]$$

Applying our two rules, that the expected value of the sum of a series of returns is equal to the sum of the expected value of each return and that the expected value of a constant times a return is equal to the constant times the expected return, we have

$$\sigma_P^2 = X_1^2 E\left[\left(R_{1j} - \bar{R}_1\right)^2\right] + 2X_1X_2E\left[\left(R_{1j} - \bar{R}_1\right)\left(R_{2j} - \bar{R}_2\right)\right] + X_2^2E\left[\left(R_{2j} - \bar{R}_2\right)^2\right]$$
$$= X_1^2\sigma_1^2 + 2X_1X_2E\left[\left(R_{1j} - \bar{R}_1\right)\left(R_{2j} - \bar{R}_2\right)\right] + X_2^2\sigma_2^2$$

where $E[(R_{1j} - \bar{R}_1)(R_{2j} - \bar{R}_2)]$ has a special name. It is called the *covariance* and will be designated as σ_{12}.[5] Substituting the symbol σ_{12} for $E[(R_{1j} - \bar{R}_1)(R_{2j} - \bar{R}_2)]$ yields

$$\sigma_P^2 = X_1^2\sigma_1^2 + X_2^2\sigma_2^2 + 2X_1X_2\sigma_{12}$$

Notice what the covariance does. It is the expected value of the product of two deviations: the deviations of the returns on security 1 from its mean $(R_{1j} - \bar{R}_1)$ and the deviations of the returns on security 2 from its mean $(R_{2j} - \bar{R}_2)$. In this sense it is very much like the variance. However, it is the product of two different deviations. As such, it can be positive or negative. It will be large when the good outcomes for each stock occur together and when the bad outcomes for each stock occur together. In this case, for good outcomes, the covariance will be the product of two large positive numbers, which is positive. When the bad outcomes occur, the covariance will be the product of two large negative numbers, which is positive. This will result in a large value for the covariance and a large variance for the portfolio. In contrast, if good outcomes for one asset are associated with bad outcomes for the other, the covariance is negative. It is negative because a positive deviation for one asset is associated with a negative deviation for the second, and the product of a positive and a negative is negative. This was what occurred when we examined a combination of assets 2 and 3.

The covariance is a measure of how returns on assets move together. Insofar as they have positive and negative deviations at similar times, the covariance is a large positive number. If they have the positive and negative deviations at dissimilar times, then the covariance is negative. If the positive and negative deviations are unrelated, it tends to be zero.

For many purposes, it is useful to standardize the covariance. Dividing the covariance between two assets by the product of the standard deviation of each asset produces a variable with the same properties as the covariance but with a range of -1 to $+1$. The measure is called the *correlation coefficient*. Letting ρ_{ik} stand for the correlation between securities i and k, we define the correlation coefficient as

$$\rho_{ik} = \frac{\sigma_{ik}}{\sigma_i\sigma_k}$$

Dividing by the product of the standard deviations does not change the properties of the covariance. It simply scales it to have values between -1 and $+1$. Let us apply these

[5]Note that when all joint outcomes are equally likely, the covariance can be expressed as

$$\sum_{j=1}^{M} \frac{\left(R_{1j} - \bar{R}_1\right)\left(R_{2j} - \bar{R}_2\right)}{M}$$

where M is the number of equally likely joint outcomes. Once again, when estimates are based on a sample of data such as actual historical returns, it is traditional to divide by $T - 1$ rather than T, where T is the number of periods in the sample.

Table 4.6 Calculating Covariances

Condition of Market	Deviations Security 1	Deviations Security 2	Product of Deviations	Deviations Security 1	Deviations Security 3	Product of Deviations
Good	(15 − 9)	(16 − 10)	36	(15 − 9)	(1 − 10)	−54
Average	(9 − 9)	(10 − 10)	0	(9 − 9)	(10 − 10)	0
Poor	(3 − 9)	(4 − 10)	36	(3 − 9)	(19 − 10)	−54
			72			−108

formulas. First, however, it is necessary to calculate covariances. Table 4.6 shows the intermediate calculations necessary to determine the covariance between securities 1 and 2 and securities 1 and 3. The sum of the deviations between securities 1 and 2 is 72. Therefore, the covariance is $72/3 = 24$ and the correlation coefficient is $24/\sqrt{24}\,\sqrt{24}$. For assets 1 and 3 the sum of the deviations is -108. The covariance is $-108/3 = -36$, and the correlation coefficient is $-36/\sqrt{24}\,\sqrt{54}$. Similar calculations can be made for all other pairs of assets, and the results are contained in Table 4.7.

Earlier we examined the results obtained by an investor with \$1.00 to spend who put \$0.60 in asset 2 and \$0.40 in asset 3. Applying the expression for variance of the portfolio, we have

$$\sigma_P^2 = \left(\frac{0.60}{1.00}\right)^2 24 + \left(\frac{0.40}{1.00}\right)^2 54 + 2\left(\frac{0.60}{1.00}\right)\left(\frac{0.40}{1.00}\right)(-36) = 0$$

This was exactly the result we obtained when we looked at the combination of the full distribution. The correlation coefficient between securities 2 and 3 is -1. This means that good and bad returns of assets 2 and 3 tended to occur at opposite times. When this situation occurs, a portfolio can always be constructed with zero risk.

Our second example was an investment in securities 1 and 4. The variance of this portfolio is

$$\sigma^2 = \left(\tfrac{1}{2}\right)^2 24 + \left(\tfrac{1}{2}\right)^2 24 = 12$$

In this case, where the correlation coefficient was zero, the risk of the portfolio was less than the risk of either of the individual securities. Once again, this is a general result. When the return patterns of two assets are independent so that the correlation coefficient and covariance are zero, a portfolio can be found that has a lower variance than either of the assets by themselves.

Table 4.7 Covariance and Correlation Coefficients (in Parentheses) between Assets

	1	2	3	4	5
1		24	−36	0	24
		(+1)	(−1)	(0)	(+1)
2			−36	0	24
			(−1)	(0)	(+1)
3				0	−36
				(0)	(−1)
4					0
					(0)
5					

As an additional check on the accuracy of the formula just derived, we calculate the variance directly. Earlier we saw there were nine possible returns when we combined assets 2 and 4. They were \$1.16, \$1.13, \$1.13, \$1.10, \$1.10, \$1.10, \$1.07, \$1.07, and \$1.04. Because we started with an investment of \$1.00, the returns are easy to determine. The returns are 16%, 13%, 13%, 10%, 10%, 10%, 7%, 7%, and 4%. By examination it is easy to see that the mean return is 10%. The deviations are 6, 3, 3, 0, 0, 0, –3, –3, and –6. The squared deviations are 36, 9, 9, 0, 0, 0, 9, 9, and 36, and the average squared deviation or variance is 108/9 = 12. This agrees with the formula developed earlier.

The final example analyzed previously was a portfolio of assets 1 and 5. In this case the variance of the portfolio is

$$\begin{aligned}\sigma_P^2 &= \left(\tfrac{1}{2}\right)^2 24 + \left(\tfrac{1}{2}\right)^2 24 + 2\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)24 \\ &= \tfrac{1}{4}(24) + \tfrac{1}{4}(24) + \tfrac{1}{2}(24) \\ &= 24\end{aligned}$$

As we demonstrated earlier, when two securities have their good and bad outcomes at the same time, the risk is not reduced by purchasing a portfolio of the two assets.

The formula for variance of a portfolio can be generalized to more than two assets. Consider first a three-asset case. Substituting the expression for return on a portfolio and the expected return of a portfolio into the general formula for variance yields

$$\begin{aligned}\sigma_P^2 &= E\left(R_P - \overline{R}_P\right)^2 \\ &= E\left[X_1R_{1j} + X_2R_{2j} + X_3R_{3j} - \left(X_1\overline{R}_1 + X_2\overline{R}_2 + X_3\overline{R}_3\right)\right]^2\end{aligned}$$

Rearranging,

$$\sigma_P^2 = E\left[X_1\left(R_{1j} - \overline{R}_1\right) + X_2\left(R_{2j} - \overline{R}_2\right) + X_3\left(R_{3j} - \overline{R}_3\right)\right]^2$$

Squaring the right-hand side yields

$$\begin{aligned}\sigma_P^2 = E\Big[&X_1^2\left(R_{1j} - \overline{R}_1\right)^2 + X_2^2\left(R_{2j} - \overline{R}_2\right)^2 + X_3^2\left(R_{3j} - \overline{R}_3\right)^2 \\ &+ 2X_1X_2\left(R_{1j} - \overline{R}_1\right)\left(R_{2j} - \overline{R}_2\right) + 2X_1X_3\left(R_{1j} - \overline{R}_1\right)\left(R_{3j} - \overline{R}_3\right) \\ &+ 2X_2X_3\left(R_{2j} - \overline{R}_2\right)\left(R_{3j} - \overline{R}_3\right)\Big]\end{aligned}$$

Applying the properties of expected return discussed earlier yields

$$\begin{aligned}\sigma_P^2 = {} & X_1^2E\left(R_{1j} - \overline{R}_1\right)^2 + X_2^2E\left(R_{2j} - \overline{R}_2\right)^2 + X_3^2E\left(R_{3j} - \overline{R}_3\right)^2 \\ & + 2X_1X_2E\left[\left(R_{1j} - \overline{R}_1\right)\left(R_{2j} - \overline{R}_2\right)\right] + 2X_1X_3E\left[\left(R_{1j} - \overline{R}_1\right)\left(R_{3j} - \overline{R}_3\right)\right] \\ & + 2X_2X_3E\left[\left(R_{2j} - \overline{R}_2\right)\left(R_{3j} - \overline{R}_3\right)\right]\end{aligned}$$

Utilizing σ_i^2 for variance of asset i and σ_{ij} for the covariance between assets i and j, we have

$$\sigma_P^2 = X_1^2\sigma_1^2 + X_2^2\sigma_2^2 + X_3^2\sigma_3^2 + 2X_1X_2\sigma_{12} + 2X_1X_3\sigma_{13} + 2X_2X_3\sigma_{23}$$

This formula can be extended to any number of assets. Examining the expression for the variance of a portfolio of three assets should indicate how. First note that the variance of each asset is multiplied by the square of the proportion invested in it. Thus the first part of the expression for the variance of a portfolio is the sum of the variances on the individual assets times the square of the proportion invested in each, or

$$\sum_{i=1}^{N}\left(X_i^2\sigma_i^2\right)$$

The second set of terms in the expression for the variance of a portfolio comprises covariance terms. Note that the covariance between each pair of assets in the portfolio enters the expression for the variance of a portfolio. With three assets the covariance between 1 and 2, 1 and 3, and 2 and 3 entered. With four assets, covariance terms between 1 and 2, 1 and 3, 1 and 4, 2 and 3, 2 and 4, and 3 and 4 would enter. Furthermore, note that each covariance term is multiplied by 2 times the product of the proportions invested in each asset. The following double summation captures the covariance terms:

$$\sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}\left(X_jX_k\sigma_{jk}\right)$$

The reader concerned that a 2 does not appear in this expression can relax. The covariance between securities 2 and 3 comes about both from $j = 2$ and $k = 3$ and from $j = 3$ and $k = 2$. This is how the term "2 times the covariance between 2 and 3" comes about. Furthermore, examining the expression for covariance shows that order does not matter; thus $\sigma_{jk} = \sigma_{kj}$. The symbol $\neq$ means k should not have the same value as j. To reemphasize the meaning of the double summation, we examine the three-security case. We have

$$\begin{aligned}\sum_{j=1}^{3}\sum_{\substack{k=1\\k\neq j}}^{3}\left(X_jX_k\sigma_{jk}\right) &= X_1X_2\sigma_{12} + X_1X_3\sigma_{13} + X_2X_1\sigma_{21}\\ &\quad + X_2X_3\sigma_{23} + X_3X_1\sigma_{31} + X_3X_2\sigma_{32}\end{aligned}$$

Because the order does not matter in calculating covariance and thus $\sigma_{12} = \sigma_{21}$, we have

$$\sum_{j=1}^{3}\sum_{\substack{k=1\\k\neq j}}^{3}\left(X_jX_k\sigma_{jk}\right) = 2X_1X_2\sigma_{12} + 2X_1X_3\sigma_{13} + 2X_2X_3\sigma_{23}$$

Putting together the variance and covariance parts of the general expression for the variance of a portfolio yields

$$\sigma_P^2 = \sum_{j=1}^{N}\left(X_j^2\sigma_j^2\right) + \sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}\left(X_jX_k\sigma_{jk}\right)$$

This formula is worth examining further. First, consider the case where all assets are independent and therefore the covariance between them is zero. This was the situation we

observed for assets 2 and 4 in our little example. In this case, $\sigma_{jk} = 0$, and the formula for variance becomes

$$\sum_{j=1}^{N}\left(X_j^2\sigma_j^2\right)$$

Furthermore, assume equal amounts are invested in each asset. With N assets, the proportion invested in each asset is $1/N$. Applying our formula yields

$$\sigma_P^2 = \sum_{j=1}^{N}(1/N)^2\sigma_j^2 = 1/N\left[\sum_{j=1}^{N}\frac{\sigma_j^2}{N}\right]$$

The term in the brackets is our expression for an average. Thus our formula reduces to $\sigma_P^2 = (1/N)\bar{\sigma}_j^2$, where $\bar{\sigma}_j^2$ represents the average variance of the stocks in the portfolio. As N gets larger and larger, the variance of the portfolio gets smaller and smaller. As N becomes extremely large, the variance of the portfolio approaches zero. This is a general result. If we have enough *independent* assets, the variance of a portfolio of these assets approaches zero.

In general, we are not so fortunate. In most markets the correlation coefficient and the covariance between assets is positive. In these markets the risk on the portfolio cannot be made to go to zero but can be much less than the variance of an individual asset. The variance of a portfolio of assets is

$$\sigma_P^2 = \sum_{j=1}^{N}\left(X_j^2\sigma_j^2\right) + \sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}\left(X_jX_k\sigma_{jk}\right)$$

Once again, consider equal investment in N assets. With equal investment, the proportion invested in any one asset X_j is $1/N$, and the formula for the variance of a portfolio becomes

$$\sigma_P^2 = \sum_{j=1}^{N}(1/N)^2\sigma_j^2 + \sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}(1/N)(1/N)\sigma_{jk}$$

Factoring out $1/N$ from the first summation and $(N-1)/N$ from the second yields

$$\sigma_P^2 = (1/N)\sum_{j=1}^{N}\left[\frac{\sigma_j^2}{N}\right] + \frac{(N-1)}{N}\sum_{j=1}^{N}\sum_{\substack{k=1\\k\neq j}}^{N}\left[\frac{\sigma_{jk}}{N(N-1)}\right]$$

Both of the terms in the brackets are averages. That the first is an average should be clear from the previous discussion. Likewise, the second term in brackets is also an average. There are N values of j and $(N-1)$ values of k. There are $N-1$ values of k because k cannot equal j so that there is one less value of k than j. In total, there are $N(N-1)$ covariance terms. Thus the second term is the summation of covariances divided by the number of covariances, and it is, therefore, an average. Replacing the summations by averages, we have

$$\sigma_P^2 = \frac{1}{N}\bar{\sigma}_j^2 + \frac{N-1}{N}\bar{\sigma}_{jk}$$

This expression is a much more realistic representation of what occurs when we invest in a portfolio of assets. The contribution to the portfolio variance of the variance of the individual securities goes to zero as N gets very large. However, the contribution of the covariance terms approaches the average covariance as N gets large. The individual risk of

Table 4.8 Effect of Diversification

Number of Securities	Expected Portfolio Variance
1	46.619
2	26.839
4	16.948
6	13.651
8	12.003
10	11.014
12	10.354
14	9.883
16	9.530
18	9.256
20	9.036
25	8.640
30	8.376
35	8.188
40	8.047
45	7.937
50	7.849
75	7.585
100	7.453
125	7.374
150	7.321
175	7.284
200	7.255
250	7.216
300	7.190
350	7.171
400	7.157
450	7.146
500	7.137
600	7.124
700	7.114
800	7.107
900	7.102
1000	7.097
Infinity	7.058

securities can be diversified away, but the contribution to the total risk caused by the covariance terms cannot be diversified away.

Table 4.8 illustrates how this relationship looks when dealing with U.S. equities. The average variance and average covariance of returns were calculated using monthly data for all stocks listed on the New York Stock Exchange (NYSE). The average variance was 46.619; the average covariance was 7.058. As more and more securities are added, the average variance on the portfolio declines until it approaches the average covariance. Rearranging the previous equation clarifies this relationship even further. Thus

$$\sigma_P^2 = (1/N)\left(\bar{\sigma}_j^2 - \bar{\sigma}_{kj}\right) + \bar{\sigma}_{kj}$$

The first term is $1/N$ times the difference between the variance of return on individual securities and the average covariance. The second term is the average covariance. This relationship clarifies the effect of diversification on portfolio risk. The minimum variance is obtained for

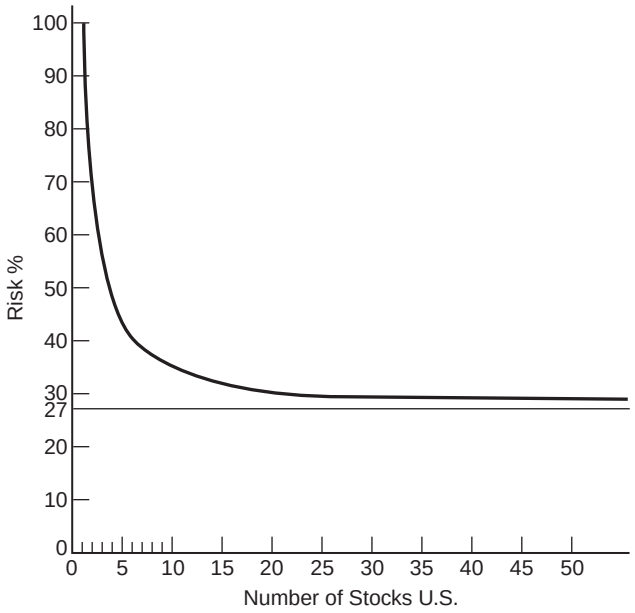


Figure 4.2 The effect of number of securities on risk of the portfolio in the United States (1975).

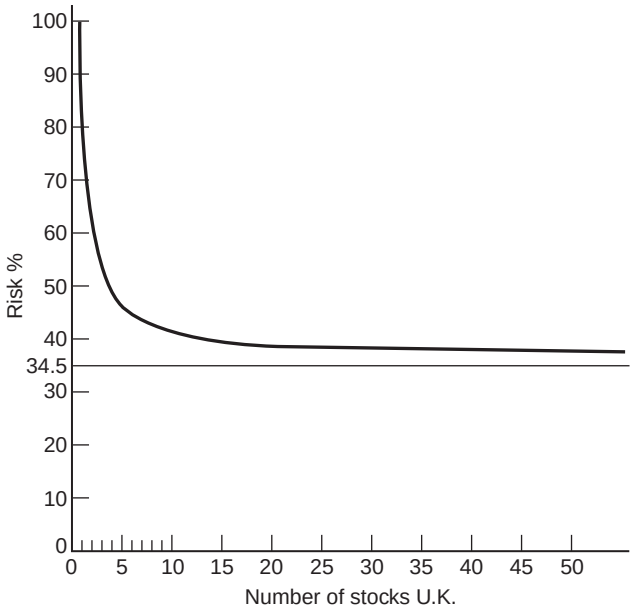


Figure 4.3 The effect of securities on risk in the United Kingdom (1975).

very large portfolios and is equal to the average covariance between all stocks in the population. As securities are added to the portfolio, the effect of the difference between the average risk on a security and the average covariance is reduced.

Figures 4.2 and 4.3 and Table 4.9 illustrate this same relationship for common equities in a number of countries. In Figure 4.2 the vertical axis is the risk of the portfolio as a percentage

Table 4.9 Percentage of the Risk on an Individual Security That Can Be Eliminated by Holding a Random Portfolio of Stocks within Selected National Markets and among National Markets (1975)

United States	73
United Kingdom	65.5
France	67.3
Germany	56.2
Italy	60.0
Belgium	80.0
Switzerland	56.0
Netherlands	76.1
International stocks	89.3

of the risk of an individual security for the United States. The horizontal axis is the number of securities in the portfolio. Figure 4.3 presents the same relationship for the United Kingdom. Table 4.9 shows the percentage of risk that can be eliminated by holding a widely diversified portfolio in each of several countries as well as with an internationally diversified portfolio. As can be seen, the effectiveness of diversification in reducing the risk of a portfolio varies from country to country. From the previous equation we know why. The average covariance relative to the variance varies from country to country. Thus, in Switzerland and Italy, securities have relatively high covariance, indicating that stocks tend to move together. Conversely, the security markets in Belgium and the Netherlands tend to have stocks with relatively low covariances. For these latter security markets, much more of the risk of holding individual securities can be diversified away. Diversification is especially useful in reducing the risk on a portfolio in these markets.

TWO CONCLUDING EXAMPLES

We close this chapter and several chapters that follow with realistic applications of the principles discussed in the chapter. These applications serve both to review the concepts presented and to demonstrate their usefulness. The two examples that follow are applications to the asset allocation decision. The first application analyzes the allocation between stocks and bonds; the second analyzes the allocation between domestic and foreign stocks.

Bond Stock Allocation

One of the major decisions facing an investor is the allocation of funds between stocks and bonds. To make this allocation, one needs to have estimates of mean returns, standard deviations of return, and either correlation coefficients or covariances. To estimate these variables, it is useful to begin by looking at historical data. Even in allocating among managed portfolios, it is useful to start by assuming that the stock and bond portfolio managers you are allocating between have performance similar to that of broad representative indexes.

The standard source for historical data is Ibbotson (2011). We have selected two indexes: a stock and a corporate bond. The stock index is a value-weighted index of the 20% of stocks on the NYSE with the largest market value. Value weighting means that the weight of the portfolio each stock represents is the market value of that stock (price times

Table 4.10 Mean Return and Standard Deviation for Combinations of Stocks and Bonds

Proportion Stocks	Proportion Bonds	Mean Return	Standard Deviation
1	0	11.8	20.3
0.9	0.1	11.26	18.29
0.8	0.2	10.72	16.33
0.7	0.3	10.18	14.43
0.6	0.4	9.64	12.63
0.5	0.5	9.1	10.98
0.4	0.6	8.56	9.56
0.3	0.7	8.02	8.47
0.2	0.8	7.48	7.85
0.1	0.9	6.94	7.83
0	1	6.4	8.40

number of shares) divided by the aggregate market value of all shares in the index. Thus the largest stocks are weighted more heavily.

The bond index is an index of corporate bond returns. The historical data are

$$\bar{R}_S = 11.8\% \qquad \sigma_S = 20.3\% \quad \rho_{SB} = .01$$
$$\bar{R}_B = 6.4\% \qquad \sigma_B = 8.4\%$$

The means and standard deviation of return for combinations of stocks and bonds varying from 100% in stocks, which is $X_S = 1$ and $X_B = 0$, to 0% in stocks are presented in Table 4.10. Note that the expected return varies linearly from 11.8% to 6.4% as we decrease the amount in the S&P and increase it in bonds. Also, the risk decreases as we put more in bonds, but not linearly. Figure 4.4 shows the various choices diagrammatically.

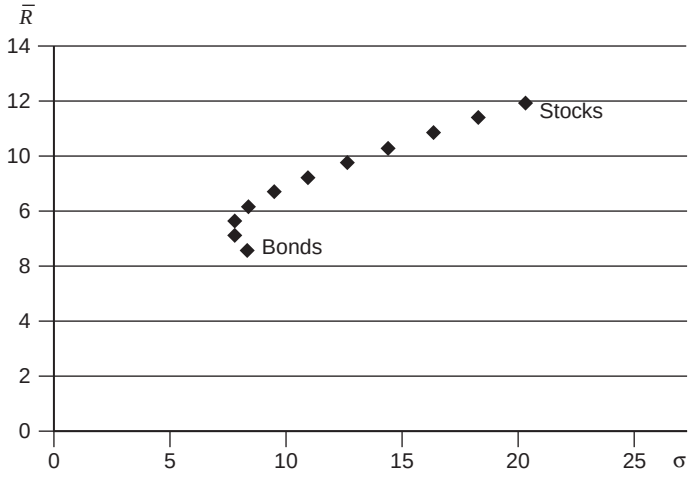


Figure 4.4 Combinations of bonds and stocks.

Table 4.11 Mean Return and Standard Deviation for Combinations of Domestic and International Stocks

Proportion Domestic	Proportion International	Mean Return	Standard Deviation
1	0	11.8	20.3
0.9	0.1	11.54	18.36
0.8	0.2	11.28	16.65
0.7	0.3	11.02	15.24
0.6	0.4	10.76	14.23
0.5	0.5	10.5	13.70
0.4	0.6	10.24	13.70
0.3	0.7	9.98	14.25
0.2	0.8	9.72	15.27
0.1	0.9	9.46	16.68
0	1	9.2	18.40

Domestic Foreign Allocation

As a second example consider the allocation decision between domestic and foreign stocks. Again we will use estimates from Ibbotson (2011). Our inputs are:

$$\overline{R}_{US} = 11.8\% \quad \sigma_{US} = 20.3\% \quad \rho_{US,INT} = .66$$
$$\overline{R}_{int} = 9.2\% \quad \sigma_{INT} = 18.4\%$$

The expected return and standard deviation of return for all combinations of the two portfolios are shown in Table 4.11 and are plotted in Figure 4.5. Note that investment in the two portfolios combined substantially reduced risk. This is a powerful demonstration of the effect of diversification.

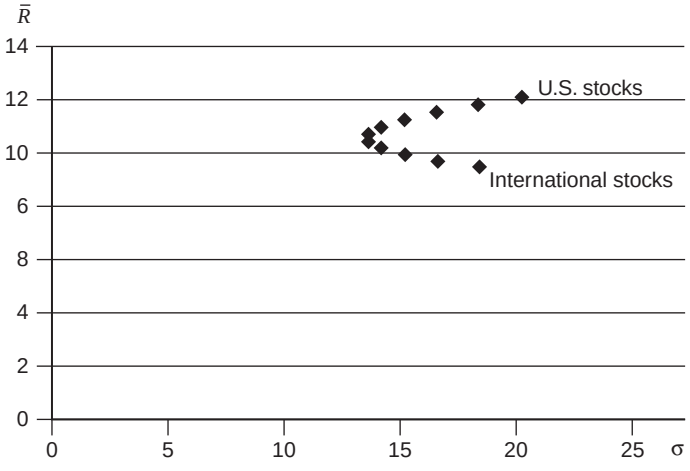


Figure 4.5 Combinations of U.S. stocks and international stocks.

CONCLUSION

In this chapter we have shown how the risk of a portfolio of assets can be very different from the risk of the individual assets composing the portfolio. This was true when we selected assets with particular characteristics, such as those shown in Table 4.3. It was also true when we simply selected assets at random, such as those shown in Tables 4.8 and 4.9.

In the following chapter we examine the relationship between the risk and the return on individual assets in more detail. We then show how the characteristics on combinations of securities can be used to define the opportunity set of investments from which the investor must make a choice. Finally, we show how the properties of these opportunities taken together with the knowledge that the investor prefers return and seeks to avoid risk can be used to define a subset of the opportunity set that will be of interest to investors.

QUESTIONS AND PROBLEMS

1. Assume that you are considering selecting assets from among the following four candidates:

Asset 1		
Market Condition	Return	Probability
Good	16	$\frac{1}{4}$
Average	12	$\frac{1}{2}$
Poor	8	$\frac{1}{4}$

Asset 2		
Market Condition	Return	Probability
Good	4	$\frac{1}{4}$
Average	6	$\frac{1}{2}$
Poor	8	$\frac{1}{4}$

Asset 3		
Market Condition	Return	Probability
Good	20	$\frac{1}{4}$
Average	14	$\frac{1}{2}$
Poor	8	$\frac{1}{4}$

Asset 4		
Rainfall	Return	Probability
Plentiful	16	$\frac{1}{3}$
Average	12	$\frac{1}{3}$
Light	8	$\frac{1}{3}$

Assume that there is no relationship between the amount of rainfall and the condition of the stock market.

A. Solve for the expected return and the standard deviation of return for each separate investment.

B. Solve for the correlation coefficient and the covariance between each pair of investments.

C. Solve for the expected return and variance of each of the portfolios shown in the following.

	Portions Invested in Each Asset			
Portfolio	Asset 1	Asset 2	Asset 3	Asset 4
a	$\frac{1}{2}$	$\frac{1}{2}$		
b	$\frac{1}{2}$		$\frac{1}{2}$	
c	$\frac{1}{2}$			$\frac{1}{2}$
d		$\frac{1}{2}$	$\frac{1}{2}$	
e			$\frac{1}{2}$	$\frac{1}{2}$
f	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	
g		$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
h	$\frac{1}{3}$		$\frac{1}{3}$	$\frac{1}{3}$
i	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

D. Plot the original assets and each of the portfolios from Part C in expected return standard deviation space.

2. Following are actual price and dividend data for three companies for each of seven months.

	Security *A*		Security *B*		Security *C*	
Time	Price	Dividend	Price	Dividend	Price	Dividend
1	$57\frac{6}{8}$		333		$106\frac{6}{8}$	
2	$59\frac{7}{8}$		368		$108\frac{2}{8}$	
3	$59\frac{3}{8}$	0.725[a]	$368\frac{4}{8}$	1.35	124	0.40
4	$55\frac{4}{8}$		$382\frac{2}{8}$		$122\frac{2}{8}$	
5	$56\frac{2}{8}$		386		$135\frac{4}{8}$	
6	59	0.725	$397\frac{6}{8}$	1.35	$141\frac{6}{8}$	0.42
7	$60\frac{2}{8}$		392		$165\frac{6}{8}$	

[a]A dividend entry on the same line as a price indicates that the return between that time period and the previous period consisted of a capital gain (or loss) and the receipt of the dividend.

A. Compute the rate of return for each company for each month.

B. Compute the average rate of return for each company.

C. Compute the standard deviation of the rate of return for each company.

D. Compute the correlation coefficient between all possible pairs of securities.

E. Compute the average return and standard deviation for the following portfolios:

$\frac{1}{2}A+\frac{1}{2}B$
$\frac{1}{2}A+\frac{1}{2}C$
$\frac{1}{2}B+\frac{1}{2}C$
$\frac{1}{3}A+\frac{1}{3}B+\frac{1}{3}C$

3. Assume that the average variance of return for an individual security is 50 and that the average covariance is 10. What is the expected variance of an equally weighted portfolio of 5, 10, 20, 50, and 100 securities?

4. In Problem 3, how many securities need to be held before the risk of a portfolio is only 10% more than minimum?
5. For the Italy data and Belgium data of Table 4.9, what is the ratio of the difference between the average variance minus average covariance and the average covariance? If the average variance of a single security is 50, what is the expected variance of a portfolio of 5, 20, and 100 securities?
6. For the data in Table 4.8, suppose an investor desires an expected variance less than 8. What is the minimum number of securities for such a portfolio?

BIBLIOGRAPHY

1. Brennan, Michael J. "The Optimal Number of Securities in a Risky Asset Portfolio When There Are Fixed Costs of Transacting: Theory and Some Empirical Results," *Journal of Financial and Quantitative Analysis*, **X,** No. 3 (Sept. 1975), pp. 483–496.
2. Elton, Edwin J., and Gruber, Martin J. "Risk Reduction and Portfolio Size: An Analytical Solution," *Journal of Business*, **50,** No. 4 (Oct. 1977), pp. 415–437.
3. ——. "Modern Portfolio Theory: 1950 to Date," *Journal of Banking and Finance*, **21,** Nos. 11–12 (Dec. 1997), pp. 1743–1759.
4. ——. "The Rationality of Asset Allocation Recommendations," *Journal of Financial and Quantitative Analysis*, **35,** No. 1 (March 2000), pp. 27–42.
5. Epps, Thomas W. "Necessary and Sufficient Conditions for the Mean-Variance Portfolio Model with Constant Risk Aversion," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 2 (June 1981), pp. 169–176.
6. Evans, L. John, and Archer, N. Stephen. "Diversification and the Reduction of Dispersion: An Empirical Analysis," *Journal of Finance*, **XXIII,** No. 5 (Dec. 1968), pp. 761–767.
7. Fisher, Lawrence, and Lorie, James. "Some Studies of Variability of Returns on Investments in Common Stocks," *Journal of Business*, **43,** No. 2 (April 1970), pp. 99–134.
8. Jennings, Edward. "An Empirical Analysis of Some Aspects of Common Stock Diversification," *Journal of Financial and Quantitative Analysis*, **VI,** No. 2 (March 1971), pp. 797–813.
9. Johnson, K., and Shannon, D. "A Note of Diversification and the Reduction of Dispersion," *Journal of Financial Economics*, **1,** No. 4 (Dec. 1974), pp. 365–372.
10. Markowitz, Harry. "Markowitz Revisited," *Financial Analysts Journal*, **32,** No. 4 (Sept.–Oct. 1976), pp. 47–52.
11. Ross, Stephen A. "Adding Risks: Samuelson's Fallacy of Large Numbers Revisited," *Journal of Financial and Quantitative Analysis*, **34,** No. 3 (Sept. 1999), pp. 323–340.
12. Rubinstein, Mark. "The Fundamental Theorem of Parameter-Preference Security Valuation," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 1 (Jan. 1973), pp. 61–69.
13. Solnick, Bruno. "The Advantages of Domestic and International Diversification," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets.* (Amsterdam: North Holland, 1975).
14. Statman, Meir. "How Many Stocks Make a Diversified Portfolio?" *Journal of Financial and Quantitative Analysis*, **22,** No. 3 (Sept. 1987), pp. 353–363.
15. Wagner, W., and Lau, S. "The Effect of Diversification on Risk," *Financial Analysts Journal*, **27,** No. 5 (Nov.–Dec. 1971), pp. 48–53.
16. Whitmore, G. A. "Diversification and the Reduction of Dispersion: A Note," *Journal of Financial and Quantitative Analysis*, **V,** No. 2 (May 1970), pp. 263–264.

5

Delineating Efficient Portfolios

In Chapter 4 we examined the return and risk characteristics of individual securities and began to study the attributes of combinations or portfolios of securities. In this chapter we look at the risk and return characteristics of combinations of securities in more detail. We start off with a reexamination of the attributes of combinations of two risky assets. In doing so, we emphasize a geometric interpretation of asset combinations. It is a short step from the analysis of the combination of two or more risky assets to the analysis of combinations of all possible risky assets. After making this step, we can delineate that subset of portfolios that will be preferred by all investors who exhibit risk avoidance and who prefer more return to less.[1] This set is usually called the efficient set or efficient frontier. Its shape will differ according to the assumptions that are made with respect to the ability of the investor to sell securities short as well as her ability to lend and borrow funds.[2] Alternative assumptions about short sales and lending and borrowing are examined.

COMBINATIONS OF TWO RISKY ASSETS REVISITED: SHORT SALES NOT ALLOWED

In Chapter 4 we began the analysis of combinations of risky assets. In this chapter we continue it. Previously, we treated the two assets as if they were individual assets, but nothing in the analysis so constrains them. In fact, when we talk about assets, we could equally well be talking about portfolios of risky assets.

Recall from Chapter 4 that the expected return on a portfolio of two assets is given by

$$\overline{R}_P = X_A\overline{R}_A + X_B\overline{R}_B \tag{5.1}$$

where

X_A is the fraction of the portfolio held in asset A

X_B is the fraction of the portfolio held in asset B

[1]In this chapter and most of those that follow, we assume that mean variance is the relevant space for portfolio analysis. See Chapter 11 for an examination of other portfolio models.

[2]Short selling is defined at a later point in this chapter.

$\bar{R}_P$ is the expected return on the portfolio
$\bar{R}_A$ is the expected return on asset A
$\bar{R}_B$ is the expected return on asset B

In addition, because we require the investor to be fully invested, the fraction she invests in A plus the fraction she invests in B must equal 1, or

$$X_A + X_B = 1$$

We can rewrite this expression as

$$X_B = 1 - X_A \tag{5.2}$$

Substituting Equation (5.2) into Equation (5.1), we can express the expected return on a portfolio of two assets as

$$\bar{R}_P = X_A \bar{R}_A + (1 - X_A)\bar{R}_B$$

Notice that the expected return on the portfolio is a simple weighted average of the expected returns on the individual securities and that the weights add to 1. The same is not necessarily true of the risk (standard deviation of the return) of the portfolio. In Chapter 4 the standard deviation of the return on the portfolio was shown to be equal to

$$\sigma_P = \left(X_A^2\sigma_A^2 + X_B^2\sigma_B^2 + 2X_A X_B \sigma_{AB}\right)^{1/2}$$

where

σ_P is the standard deviation of the return on the portfolio
σ_A^2 is the variance of the return on security A
σ_B^2 is the variance of the return on security B
σ_{AB} is the covariance between the returns on security A and security B

If we substitute Equation (5.2) into this expression, we obtain

$$\sigma_P = \left[X_A^2\sigma_A^2 + (1 - X_A)^2\sigma_B^2 + 2X_A(1 - X_A)\sigma_{AB}\right]^{1/2} \tag{5.3}$$

Recalling that $\sigma_{AB} = \rho_{AB}\sigma_A\sigma_B$, where ρ_{AB} is the correlation coefficient between securities A and B, Equation (5.3) becomes

$$\sigma_P = \left[X_A^2\sigma_A^2 + (1 - X_A)^2\sigma_B^2 + 2X_A(1 - X_A)\rho_{AB}\sigma_A\sigma_B\right]^{1/2} \tag{5.4}$$

The standard deviation of the portfolio is not, in general, a simple weighted average of the standard deviation of each security. Cross-product terms are involved, and the weights do not, in general, add to 1. To learn more about this relationship, we now study some specific cases involving different degrees of comovement between securities.

We know that a correlation coefficient has a maximum value of $+1$ and a minimum value of -1. A value of $+1$ means that two securities will always move in perfect unison, whereas a value of -1 means that their movements are exactly opposite to each other. We start with an examination of these extreme cases, then we turn to an examination of some intermediate values for the correlation coefficients. As an aid in interpreting results, we examine a specific example as well as general expressions for risk and return. For the example, we consider two stocks: a large manufacturer of automobiles ("Colonel Motors") and an

electric utility company operating in a large eastern city ("Separated Edison"). Assume the stocks have the following characteristics:

	Expected Return	Standard Deviation
Colonel Motors (*C*)	14%	6%
Separated Edison (*S*)	8%	3%

As you might suspect, the car manufacturer has a bigger expected return and a bigger risk than the electric utility.

Case 1—Perfect Positive Correlation ($\rho = +1$)

Let the subscript *C* stand for Colonel Motors and the subscript *S* stand for Separated Edison. If the correlation coefficient is +1, then the equation for the risk on the portfolio, Equation (5.4), simplifies to

$$\sigma_P = \left[X_C^2\sigma_C^2 + (1-X_C)^2\sigma_S^2 + 2X_C(1-X_C)\sigma_C\sigma_S\right]^{1/2} \tag{5.5}$$

Note that the term in square brackets has the form $X^2 + 2XY + Y^2$ and thus can be written as

$$\left[X_C\sigma_C + (1-X_C)\sigma_S\right]^2$$

Because the standard deviation of the portfolio is equal to the positive square root of this expression, we know that

$$\sigma_P = X_C\sigma_C + (1-X_C)\sigma_S$$

while the expected return on the portfolio is

$$\bar{R}_P = X_C\bar{R}_C + (1-X_C)\bar{R}_S$$

Thus with the correlation coefficient equal to +1, both risk and return of the portfolio are simply linear combinations of the risk and return of each security. In footnote 3 we show that the form of these two equations means that all combinations of two securities that are perfectly correlated will lie on a straight line in risk and return space.[3] We now illustrate that this is true for the stocks in our example. For the two stocks under study,

[3]Solving for X_C in the expression for standard deviation yields

$$X_C = \frac{\sigma_P - \sigma_S}{\sigma_C - \sigma_S}$$

Substituting this into the expression for expected return yields

$$\bar{R}_P = \frac{\sigma_P - \sigma_S}{\sigma_C - \sigma_S}\bar{R}_C + \left(1 - \frac{\sigma_P - \sigma_S}{\sigma_C - \sigma_S}\right)\bar{R}_S$$

$$\bar{R}_P = \left(\bar{R}_S - \frac{\bar{R}_C - \bar{R}_S}{\sigma_C - \sigma_S}\sigma_S\right) + \left(\frac{\bar{R}_C - \bar{R}_S}{\sigma_C - \sigma_S}\right)\sigma_P$$

which is the equation of a straight line connecting security *C* and security *S* in expected return standard deviation space.

Table 5.1 The Expected Return and Standard Deviation of a Portfolio of Colonel Motors and Separated Edison When $\rho = +1$

X_C	0	0.2	0.4	0.5	0.6	0.8	1.0
$\bar{R}_P$	8.0	9.2	10.4	11	11.6	12.8	14.0
σ_P	3.0	3.6	4.2	4.5	4.8	5.4	6.0

$$\bar{R}_P = \frac{\sigma_P - \sigma_S}{\sigma_C - \sigma_S}\bar{R}_C + \left(1 - \frac{\sigma_P - \sigma_S}{\sigma_C - \sigma_S}\right)\bar{R}_S$$

Table 5.1 presents the return on a portfolio for selected values of X_C, and Figure 5.1 presents a graph of this relationship. Note that the relationship is a straight line. The equation of the straight line could easily be derived as follows. Utilizing the equation presented earlier for σ_P to solve for X_C yields

$$X_C = \frac{\sigma_P}{3} - 1$$

Substituting this expression for X_C into the equation for $\bar{R}_P$ and rearranging yields[4]

$$\bar{R}_P = 2 + 2\sigma_P$$

In the case of perfectly correlated assets, the return and risk on the portfolio of the two assets is a weighted average of the return and risk on the individual assets. There is no reduction in risk from purchasing both assets. This can be seen by examining Figure 5.1

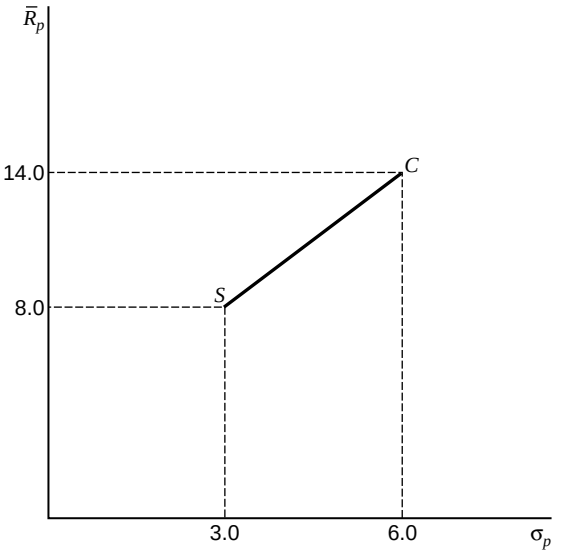


Figure 5.1 Relationship between expected return and standard deviation when $\rho = +1$.

[4]An alternative way to derive this equation is to substitute the appropriate values for the two firms into the equation derived in footnote 3. This yields

$$\bar{R}_P = 8 + 6\frac{(\sigma_P - 3)}{3} = 2 + 2\sigma_P$$

and noting that combinations of the two assets lie along a straight line connecting the two assets. Nothing has been gained by diversifying rather than purchasing the individual assets.

Case 2—Perfect Negative Correlation ($\rho = -1.0$)

We now examine the other extreme: two assets that move perfectly together but in exactly opposite directions. In this case the standard deviation of the portfolio is [from Equation (5.4) with $\rho = -1.0$]

$$\sigma_P = \left[X_C^2\sigma_C^2 + (1-X_C)^2\sigma_S^2 - 2X_C(1-X_C)\sigma_C\sigma_S\right]^{1/2} \tag{5.6}$$

Once again, the equation for standard deviation can be simplified. The term in the brackets is equivalent to either of the following two expressions:

$$\left[X_C\sigma_C - (1-X_C)\sigma_S\right]^2$$

or

$$\left[-X_C\sigma_C + (1-X_C)\sigma_S\right]^2 \tag{5.7}$$

Thus σ_P is either

$$\sigma_P = X_C\sigma_C - (1-X_C)\sigma_S$$

or

$$\sigma_P = -X_C\sigma_C + (1-X_C)\sigma_S \tag{5.8}$$

Because we took the square root to obtain an expression for σ_P and because the square root of a negative number is imaginary, either of the preceding equations holds only when its right-hand side is positive. A further examination shows that the right-hand side of one equation is simply -1 times the other. Thus each equation is valid only when the right-hand side is positive. Because one is always positive when the other is negative (except when both equations equal zero), there is a unique solution for the return and risk of any combination of securities *C* and *S*. These equations are very similar to the ones we obtained when we had a correlation of $+1$. Each also plots as a straight line when σ_P is plotted against X_C. Thus one would suspect that an examination of the return on the portfolio of two assets as a function of the standard deviation would yield two straight lines, one for each expression for σ_P. As we observe in a moment, this is, in fact, the case.[5]

The value of σ_P for Equation (5.7) or (5.8) is always smaller than the value of σ_P for the case where $\rho = +1$ [Equation (5.5)] for all values of X_C between 0 and 1. Thus the risk on a portfolio of assets is always smaller when the correlation coefficient is -1 than when it is $+1$. We can go one step further. If two securities are perfectly negatively correlated (i.e., they move in exactly opposite directions), it should always be possible to find some combination of these two securities that has zero risk. By setting either Equation (5.7) or (5.8) equal to 0, we find that a portfolio with $X_C = \sigma_S/(\sigma_S + \sigma_C)$ will have zero risk. Because $\sigma_S > 0$ and $\sigma_S + \sigma_C > \sigma_S$, this implies that $0 < X_C < 1$ or that the zero-risk portfolio will always involve positive investment in both securities.

[5]This occurs for the same reason that the analysis for $\rho = +1$ led to one straight line, and the mathematical proof is analogous to that presented for the case of $\rho = +1$.

Now let us return to our example. Minimum risk occurs when $X_C = 3/(3 + 6) = \frac{1}{3}$. Furthermore, for the case of perfect negative correlation,

$$\bar{R}_P = 8 + 6X_C$$
$$\sigma_P = 6X_C - 3(1 - X_C)$$

or

$$\sigma_P = -6X_C + 3(1 - X_C)$$

there are two equations relating σ_P to X_C. Only one is appropriate for any value of X_C. The appropriate equation to define σ_P for any value of X_C is that equation for which $\sigma_P \geq 0$. Note that if $\sigma_P > 0$ from one equation, then $\sigma_P < 0$ for the other. Table 5.2 presents the return on the portfolio for selected values of X_C, and Figure 5.2 presents a graph of this relationship.[6]

Notice that a combination of the two securities exists that provides a portfolio with zero risk. Employing the formula developed before for the composition of the zero-risk portfolio, we find that X_C should equal $3/(3 + 6)$ or $\frac{1}{3}$. We can see this is correct from Figure 5.2 or by substituting $\frac{1}{3}$ for X_C in the equation for portfolio risk given previously. We have once again demonstrated the most powerful result of diversification: the ability of combinations of securities to reduce risk. In fact, it is not uncommon for combinations of two securities to have less risk than either of the assets in the combination.

We have now examined combinations of risky assets for perfect positive and perfect negative correlation. In Figure 5.3 we have plotted both of these relationships on the same graph. From this graph we should be able to see intuitively where portfolios of these two stocks should lie if correlation coefficients took on intermediate values. From the expression for the standard deviation [Equation (5.4)], we see that for any value for X_C between 0 and 1, the lower the correlation, the lower the standard deviation of the portfolio. The standard deviation reaches its lowest value for $\rho = -1$ (curve *SBC*) and its highest value for $\rho = +1$ (curve *SAC*). Therefore these two curves should represent the limits within which all portfolios of these two securities must lie for intermediate values of the correlation coefficient. We would speculate that an intermediate correlation might produce a curve such as *SOC* in Figure 5.3. We demonstrate this by returning to our example and constructing the relationship between risk and return for portfolios of our two securities when the correlation coefficient is assumed to be 0 and $+0.5$.

Table 5.2 The Expected Return and Standard Deviation of a Portfolio of Colonel Motors and Separated Edison When $\rho = -1$

X_C	0	0.2	0.4	0.6	0.8	1.0
$\bar{R}_P$	8.0	9.2	10.4	11.6	12.8	14.0
σ_P	3.0	1.2	0.6	2.4	4.2	6.0

[6]The equation for $\bar{R}_P$ as a function of σ_P can be obtained by solving the expression relating σ_P and X_C for X_C and using this to eliminate X_C in the expression for $\bar{R}_P$. This yields

$$\bar{R}_P = 8 + 6\left(\frac{\sigma_P + 3}{6 + 3}\right) = 10 + \frac{2}{3}\sigma_P$$

or

$$\bar{R}_P = 8 + 6\left(\frac{\sigma_P - 3}{-6 - 3}\right) = 10 - \frac{2}{3}\sigma_P$$

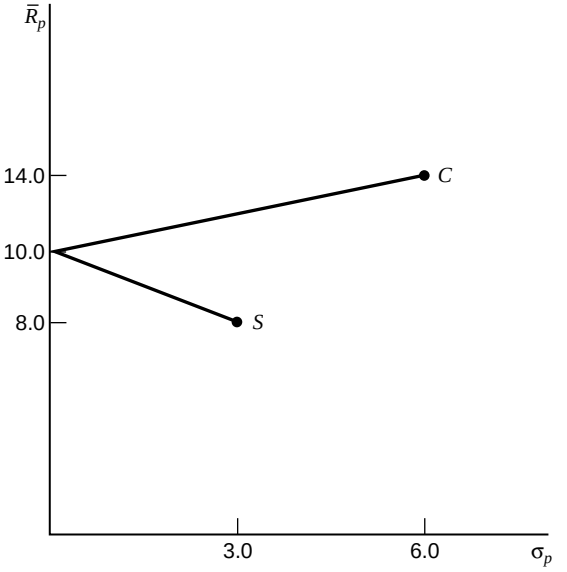


Figure 5.2 Relationship between expected return and standard deviation when $\rho = -1$.

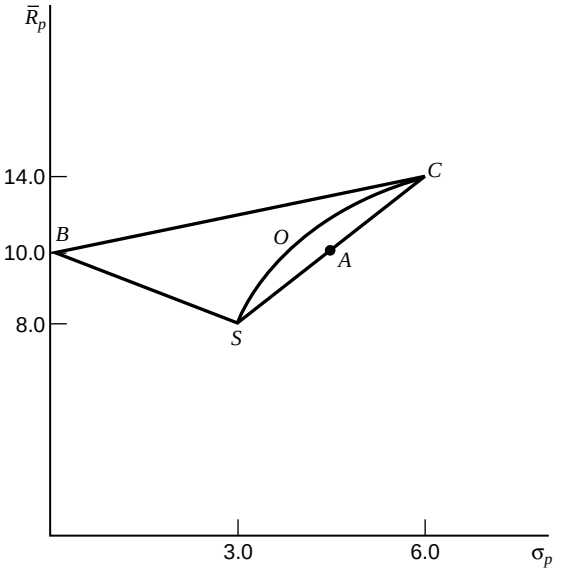


Figure 5.3 Relationship between expected return and standard deviation for various correlation coefficients.

Case 3—No Relationship between Returns on the Assets ($\rho = 0$)

The expression for return on the portfolio remains unchanged; however, because the covariance term drops out, the expression for standard deviation becomes

$$\sigma_P = \left[X_C^2 \sigma_C^2 + (1 - X_C)^2 \sigma_S^2 \right]^{1/2}$$

Table 5.3 The Expected Return and Standard Deviation for a Portfolio of Colonel Motors and Separated Edison When $\rho = 0$

X_C	0	0.2	0.4	0.6	0.8	1.0
$\bar{R}_P$	8.0	9.2	10.4	11.6	12.8	14.0
σ_P	3.00	2.68	3.00	3.79	4.84	6.0

For our example, this yields

$$\sigma_P = \left[(6)^2 X_C^2 + (3)^2 (1 - X_C)^2\right]^{1/2}$$

$$\sigma_P = \left[45X_C^2 - 18X_C + 9\right]^{1/2}$$

Table 5.3 presents the returns and standard deviation on the portfolio of Colonel Motors and Separated Edison for selected values of X_C.

A graphical presentation of the risk and return on these portfolios is shown in Figure 5.4. One point on this figure is worth special attention: the portfolio that has minimum risk. This portfolio can be found in general by looking at the equation for risk:

$$\sigma_P = \left[X_C^2\sigma_C^2 + (1 - X_C)^2\sigma_S^2 + 2X_C(1 - X_C)\sigma_C\sigma_S\rho_{CS}\right]^{1/2}$$

To find the value of X_C that minimizes this equation, we take the derivative of it with respect to X_C, set the derivative equal to zero, and solve for X_C. The derivative is

$$\frac{\partial\sigma_P}{\partial X_C} = \left(\frac{1}{2}\right)\frac{\left[2X_C\sigma_C^2 - 2\sigma_S^2 + 2X_C\sigma_S^2 + 2\sigma_C\sigma_S\rho_{CS} - 4X_C\sigma_C\sigma_S\rho_{CS}\right]}{\left[X_C^2\sigma_C^2 + (1 - X_C)^2\sigma_S^2 + 2X_C(1 - X_C)\sigma_C\sigma_S\rho_{CS}\right]^{1/2}}$$

Setting this equal to zero and solving for X_C yields

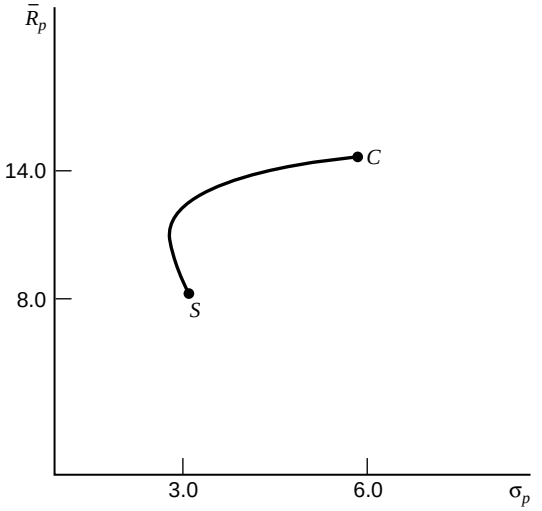


Figure 5.4 Relationship between expected return and standard deviation when $\rho = 0$.

$$X_C = \frac{\sigma_S^2 - \sigma_C \sigma_S \rho_{CS}}{\sigma_C^2 + \sigma_S^2 - 2\sigma_C \sigma_S \rho_{CS}} \tag{5.9}$$

In the present case ($\rho_{CS} = 0$), this reduces to

$$X_C = \frac{\sigma_S^2}{\sigma_C^2 + \sigma_S^2}$$

Continuing with the previous example, we find that the value of X_C that minimizes risk is

$$X_C = \frac{9}{9+36} = \frac{1}{5} = 0.20$$

This is the minimum-risk portfolio that was shown in Figure 5.4.

Case 4—Intermediate Risk ($\rho = 0.5$)

The correlation between any two actual stocks is almost always greater than 0 and considerably less than 1. To show a more typical relationship between risk and return for two stocks, we have chosen to examine the relationship when $\rho = +0.5$.

The equation for the risk of portfolios composed of Colonel Motors and Separated Edison when the correlation is 0.5 is

$$\sigma_P = \left[(6)^2 X_C^2 + (3)^2 (1 - X_C)^2 + 2X_C(1 - X_C)(3)(6)\left(\tfrac{1}{2}\right)\right]^{1/2}$$

$$\sigma_P = \left(27X_C^2 + 9\right)^{1/2}$$

Table 5.4 presents the returns and risks on alternative portfolios of our two stocks when the correlation between them is 0.5.

This risk–return relationship is plotted in Figure 5.5 along with the risk–return relationships for other intermediate values of the correlation coefficient. Notice that in this example, if $\rho = 0.5$, then the minimum risk is obtained at a value of $X_C = 0$ or where the investor has placed 100% of his funds in Separated Edison. This point could have been derived analytically from Equation (5.9). Employing this equation yields

$$X_C = \frac{9 - 18(0.5)}{9 + 36 - 2(18)(0.5)} = 0$$

In this example (i.e., $\rho_{CS} = 0.5$), there is no combination of the two securities that is less risky than the least risky asset by itself, though combinations are still less risky than they were in the case of perfect positive correlation. The particular value of the correlation coefficient for which no combination of two securities is less risky than the least risky security depends on the characteristics of the assets in question. Specifically, for all assets, there is

Table 5.4 The Expected Return and Standard Deviation of a Portfolio of Colonel Motors and Separated Edison When $\rho = 0.5$

X_C	0	0.2	0.4	0.6	0.8	1.0
$\bar{R}_P$	8.0	9.2	10.4	11.6	12.8	14.0
σ_P	3.00	3.17	3.65	4.33	5.13	6.00

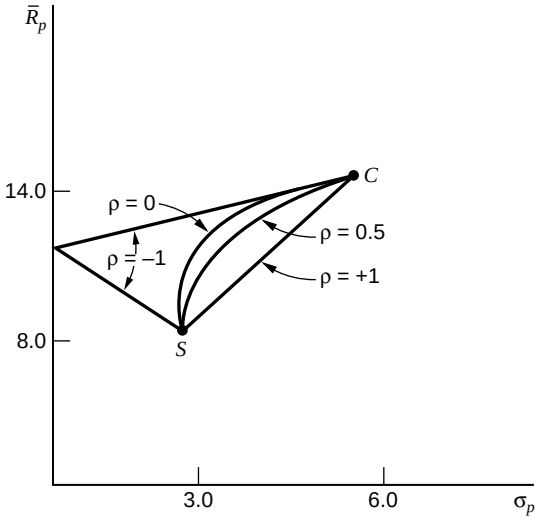


Figure 5.5 Relationship between expected return and standard deviation of return for various correlation coefficients.

some value of ρ such that the risk of the portfolio can no longer be made less than the risk of the least risky asset in the portfolio.[7]

We have developed some insights into combinations of two securities or portfolios from the analysis performed to this point. First, we have noted that the lower (closer to -1.0) the correlation coefficient between assets, all other attributes held constant, the higher the payoff from diversification. Second, we have seen that combinations of two assets can never have more risk than that found on a straight line connecting the two assets in expected return standard deviation space. Finally, we have produced a simple expression for finding the minimum variance portfolio when two assets are combined in a portfolio. We can use this to gain more insight into the shape of the curve along which all possible combinations of assets must lie in expected return standard deviation space. This curve, which is called the portfolio possibilities curve, is the subject of the next section.

THE SHAPE OF THE PORTFOLIO POSSIBILITIES CURVE

Reexamine the earlier figures in this chapter and note that the portion of the portfolio possibility curve that lies above the minimum variance portfolio is concave, whereas that which lies below the minimum variance portfolio is convex.[8] This is not due to the peculiarities of the examples we have chosen but rather is a general characteristic of all portfolio problems.

[7]The value of the correlation coefficient where this occurs is easy to determine. Equation (5.9) is the expression for the fraction of the portfolio to be held in X_C to minimize risk. Assume X_S is the least risky asset. When X_C equals zero in Equation (5.9), that means that 100% of the funds are invested in the least risky asset (i.e., X_S equals 1) to obtain the least risky portfolio. Setting X_C equal to zero in Equation (5.9) and solving for ρ_{CS} gives $\rho_{CS} = \sigma_S/\sigma_C$. So when ρ_{CS} is equal to σ_S/σ_C, X_C will equal zero, and the least risky "combination" of assets will be 100% invested in the least risky asset alone. If ρ_{CS} is greater than σ_S/σ_C, then the least risky combination involves short selling *C*.

[8]A concave curve is one where a straight line connecting any two points on the curve lies entirely under the curve. If a curve is convex, a straight line connecting any two points lies totally above the curve. The only exception to this is that a straight line is both convex and concave and so can be referred to as either.

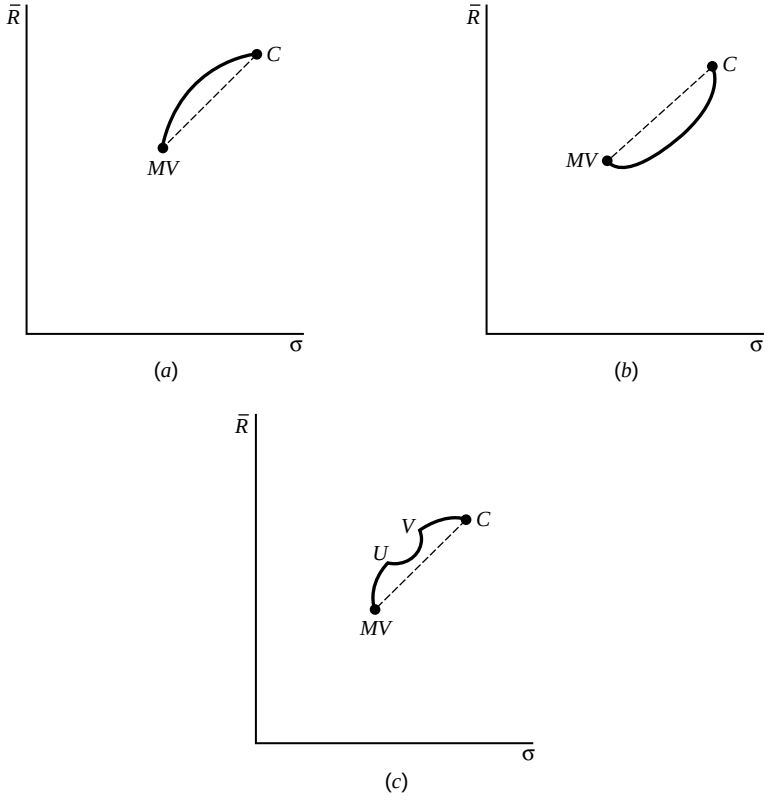


Figure 5.6 Various possible relationships for expected return and standard deviation when the minimum variance portfolio and Colonel Motors are combined.

This can easily be demonstrated. Remember that the equations and diagrams we have developed are appropriate for all combinations of securities and portfolios. We now examine combinations of the minimum variance portfolio and an asset that has a higher return and risk.

Figures 5.6*a*, 5.6*b*, and 5.6*c* represent three hypothesized shapes for combinations of Colonel Motors and the minimum variance portfolio. The shape depicted in 5.6*b* cannot be possible because we have demonstrated that combinations of assets cannot have more risk than that found on a straight line connecting two assets (and that occurs only in the case of perfect positive correlation). But what about the shape presented in Figure 5.6*c*? Here all portfolios have less risk than the straight line connecting Colonel Motors and the minimum variance portfolio. However, this is impossible. Examine the portfolios labeled *U* and *V.* These are simply combinations of the minimum variance portfolio and Colonel Motors. Since *U* and *V* are portfolios, all combinations of *U* and *V* must lie either on a straight line connecting *U* and *V* or above such a straight line.[9] Hence 5.6*c* is impossible, and the only legitimate shape is that shown in 5.6*a*, which is a concave curve. Analogous reasoning can be used to show that if we consider combinations of the minimum variance portfolio and a security or portfolio with higher variance and lower return, the curve must be convex, that is, it must look like Figure 5.7*a* rather than 5.7*b* or 5.7*c*.

[9]If the correlation between *U* and *V* equals +1, they will be on the straight line. If it is less than +1, the risk must be less, so combinations must be above the straight line.

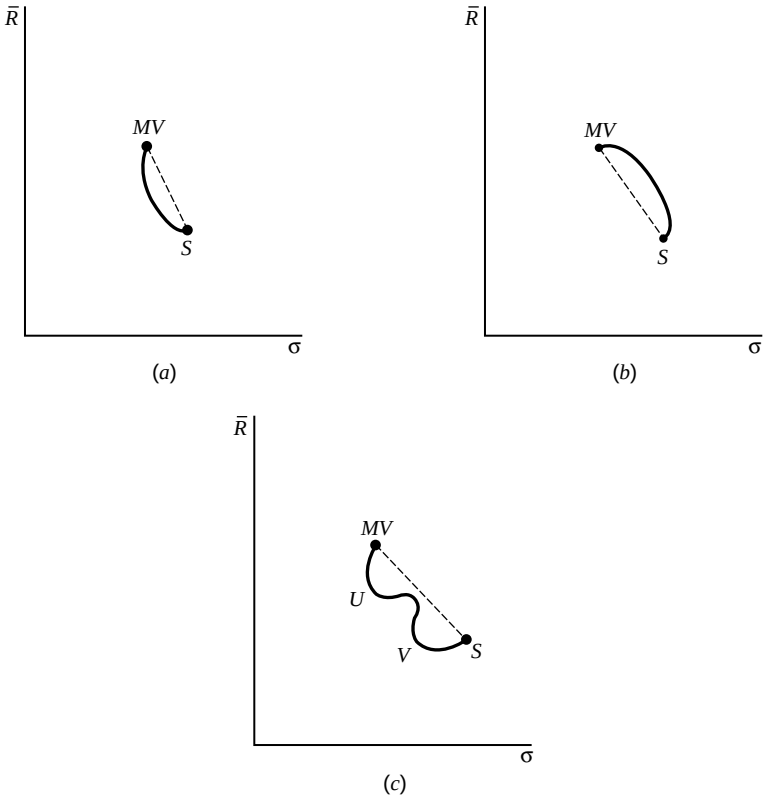


Figure 5.7 Various possible relationships between expected return and standard deviation of return when the minimum variance portfolio is combined with portfolio *S*.

Now that we understand the risk–return properties of combinations of two assets, we are in a position to study the attributes of combinations of all risky assets.

The Efficient Frontier with No Short Sales

In theory we could plot all conceivable risky assets and combinations of risky assets in a diagram in return standard deviation space. We used the words "in theory" not because there is a problem in calculating the risk and return on a stock or portfolio but because there are an infinite number of possibilities that must be considered. Not only must all possible groupings of risky assets be considered but all groupings must be considered in all possible percentage compositions.

If we were to plot all possibilities in risk–return space, we would get a diagram like Figure 5.8. We have taken the liberty of representing combinations as a finite number of points in constructing the diagram. Let us examine the diagram and see if we can eliminate any part of it from consideration by the investor. In Chapter 4 we reasoned that an investor would prefer more return to less and would prefer less risk to more. Thus, if we could find a set of portfolios that

1. offered a bigger return for the same risk or
2. offered a lower risk for the same return

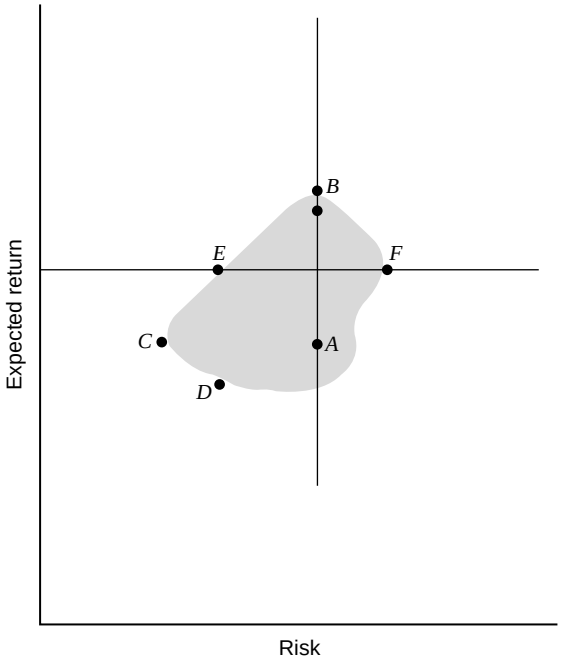


Figure 5.8 Risk and return possibilities for various assets and portfolios.

we would have identified all portfolios an investor could consider holding. All other portfolios could be ignored.

Let us take a look at Figure 5.8. Examine portfolios *A* and *B.* Note that portfolio *B* would be preferred by all investors to portfolio *A* because it offers a higher return with the same level of risk. We can also see that portfolio *C* would be preferable to portfolio *A* because it offers less risk at the same level of return. Notice that, at this point in our analysis, we can find no portfolio that dominates portfolio *C* or portfolio *B.* It should be obvious at this point that an efficient set of portfolios cannot include interior portfolios. We can reduce the possibility set even further. For any point in risk–return space, we want to move as far as possible in the direction of increasing return and as far as possible in the direction of decreasing risk. Examine point *D*, which is an exterior point. We can eliminate *D* from further consideration given the existence of portfolio *E*, which has more return for the same risk. This is true for every other portfolio as we move up the outer shell from point *D* to point *C.* Point *C* cannot be eliminated because there is no portfolio that has less risk for the same return or more return for the same risk. But what is point *C*? It is the global minimum variance portfolio.[10] Now examine point *F.* Point *F* is on the outer shell, but point *E* has less risk for the same return. As we move up the outer shell curve from point *F*, all portfolios are dominated until we come to portfolio *B.* Portfolio *B* cannot be eliminated, for there is no portfolio that has the same return and less risk or the same risk and more return than point *B.* Point *B* represents that portfolio (usually a single security) that offers the highest expected return of all portfolios. Thus the efficient set consists of the envelope curve of all portfolios that lie between the global minimum variance portfolio and the maximum return portfolio. This set of portfolios is called the *efficient frontier*.

[10]The global minimum variance portfolio is that portfolio that has the lowest risk of any feasible portfolio.

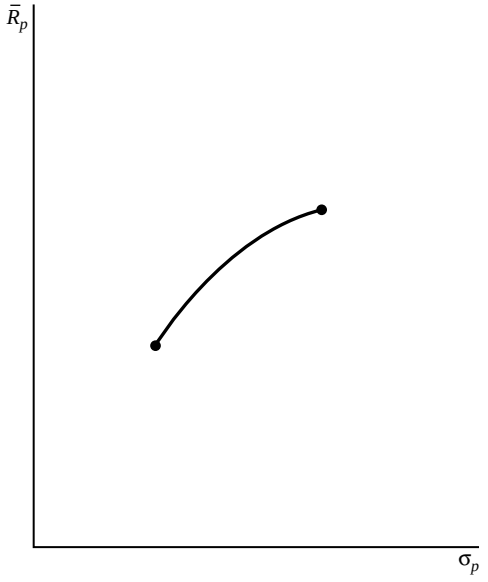


Figure 5.9 The efficient frontier.

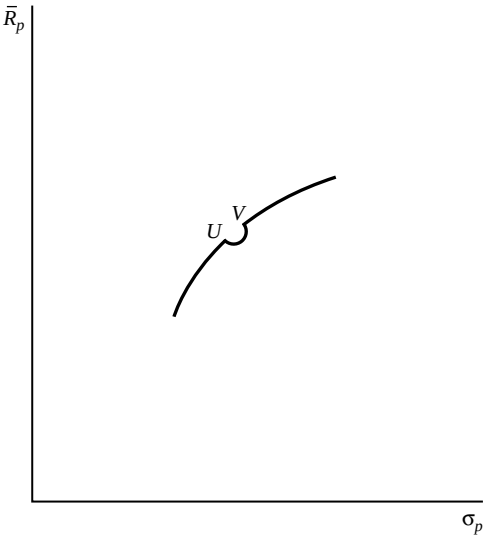


Figure 5.10 An impossible shape for the efficient frontier.

Figure 5.9 represents a graph of the efficient frontier. Notice that we have drawn the efficient frontier as a concave function. The proof that it must be concave follows logically from the earlier analysis of the combination of two securities or portfolios. The efficient frontier cannot contain a convex region such as that shown in Figure 5.10 because, as argued earlier, *U* and *V* are portfolios, and combinations of two portfolios must be concave.[11]

[11]Furthermore, there can be linear segments if the two efficient portfolios are perfectly correlated. Because a linear relationship is both concave and convex, we can still refer to the efficient frontier as concave.

Up to this point, we have seen that the efficient frontier is a concave function in expected return standard deviation space that extends from the minimum variance portfolio to the maximum return portfolio. The portfolio problem, then, is to find all portfolios along this frontier. The computational procedures necessary to do so will be examined in Chapter 6.

The Efficient Frontier with Short Sales Allowed

In the stock market (and many other capital markets), an investor can often sell a security that she does not own. This process is called short selling and is described in Chapter 3; however, the mechanics of short selling are worth repeating here. It involves in essence taking a negative position in a security. Short sales exist in sizable amounts on the New York Stock Exchange (as well as other securities markets) and the number of short sales in New York Stock Exchange stocks is reported in the *New York Times* every Monday. In a moment we will discuss the incorporation of short sales into our analysis. Before we do so, however, it is worthwhile pointing out that we have not been wasting our time by studying the case where short sales are disallowed. There are two reasons why this is true. The first is that most institutional investors do not short sell. Many institutions are forbidden by law from short selling, whereas still others operate under a self-imposed constraint forbidding short sales. The second is that the incorporation of short sales into our analysis involves only a minor extension of the analysis we have developed up to this point.

In this section we employ a simplified description of the way short sales work. This has been the general description of short sales in the literature, but in footnotes and in Chapter 6, we present both the deficiencies of this description and an alternative, more realistic description of short sales. Our description of short sales, which treats short sales as the ability to sell a security without owning it, assumes that there are no special transaction costs involved in this process. Let us see how this process might work.

Let us assume an investor believes that the stock of ABC company, which currently sells for \$100 per share, is likely to be selling for \$95 per share (expected value) at the end of the year. In addition, the investor expects ABC company to pay a \$3.00 dividend at the end of the year. If the investor were to buy one share of ABC stock, the cash flow would be −\$100.00 at time zero, when the stock is purchased, and +\$3.00 from the dividend, plus +\$95.00 from selling the stock at time 1. The cash flows are

	Time	
	0	1
Purchase stock	−100	
Dividend		+ 3
Sell stock		+95
Total cash flow	−100	+98

Unless this stock had very unusual correlations with other securities, it is unlikely that an investor with these expectations would want to hold any of it in his own portfolio. In fact, an investor would really like to own negative amounts of it. How might the investor do so? Assume a friend, Joelle, owned a share of ABC company and that the friend had different expectations and wished to continue holding it. The investor might borrow Joelle's stock under the promise that she will be no worse off lending him the stock. The investor could then sell the stock, receiving \$100. When the company pays the \$3.00 dividend, the investor must reach into his own pocket and pay Joelle \$3.00. He has a cash flow of −\$3.00. He has to do this because neither he nor Joelle now owns the stock, and he

promised that Joelle would be no worse off by lending him the stock. Now, at the end of the year, the investor could purchase the stock for $95 and give it back to Joelle. The cash flows for the investor are

	Time	
	0	1
Sell stock	+100	
Pay dividend		− 3
Buy stock		−95
Total cash flow	+100	−98

Notice in the example that the lender of the stock is no worse off by the process and that the borrower has been able to create a security that has the opposite characteristics of buying a share of the ABC company. In the real world, Joelle might require some added compensation for lending her stock, but we will continue to use this simplified description of short selling in analyzing portfolio possibilities.[12]

It was clear that when an investor expected the return on a security to be negative, short sales made sense. Even in the case where returns are positive, short sales can make sense, for the cash flow received at time zero from short selling one security can be used to purchase a security with a higher expected return. Return to an example employing Colonel Motors and Separated Edison. Recall that the expected return for Separated Edison was 8%, whereas it was 14% for Colonel Motors. If we disallow short sales, the highest return an investor can get is 14%, by placing 100% of the funds in Colonel Motors. With short sales, higher returns can be earned by short selling Separated Edison and placing the investor's original capital plus the initial cash flow from short sales in Colonel Motors. In doing so, however, there is a commensurate increase in risk. To see this more formally, we return to the case where the correlation coefficient between the two securities is assumed to be 0.5 and see what happens when we allow short sales. The earlier calculations in Table 5.4 and the diagram in Figure 5.5 are still valid, but now they must be extended to consider values of X greater than 1 and less than 0. Some sample calculations are shown in Table 5.5.

Table 5.5 The Expected Return and Standard Deviation When Short Sales Are Allowed

X_C	−1	−0.8	−0.6	−0.4	−0.2	+1.2	+1.4	+1.6	+1.8	+2.0
$\bar{R}$	2.0	3.2	4.4	5.6	6.8	15.2	16.4	17.6	18.8	20.0
σ	6.0	5.13	4.33	3.65	3.17	6.92	7.87	8.84	9.82	10.82

The new diagram with short sales is shown in Figure 5.11. The reader should note that with short sales, portfolios exist that give infinite expected rates of return. This should not be too surprising, because with short sales, one can sell securities with low expected returns and use the proceeds to buy securities with high expected returns. For example, suppose an investor had $100 to invest in Colonel Motors and Separated Edison. The investor could

[12]In the case of actual short sales, a broker plays the role of the friend and demands that funds be put up as security for the loan of the stock. These funds are in addition to the proceeds from the short sale. Because, in most cases, the amount of the funds that must be put up is quite large and the broker pays no return on these funds, the description of short sales commonly used in the literature overstates the return from short sales.

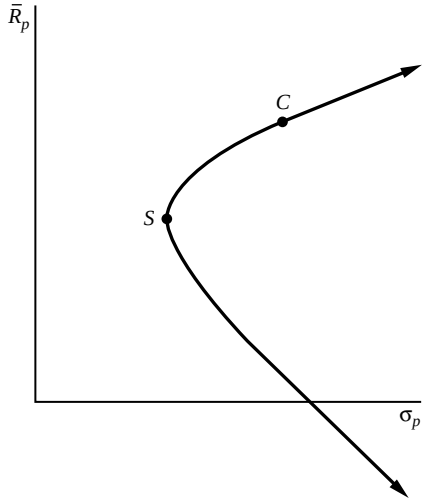


Figure 5.11 Expected return standard deviation combinations of Colonel Motors and Separated Edison when short sales are allowed.

place the entire $100 in Colonel Motors and get a return of $14, or 14%. Conversely, the investor could sell $1,000 worth of Separated Edison stock short and buy $1,100 worth of Colonel Motors. The expected earnings on the investment in Colonel Motors is $154, whereas the expected cost of borrowing Separated Edison is $80. Therefore, the expected return would be $74, or 74%, on the original $100 investment. Is this a preferred position? The expected return would increase from 14% to 74%, but the standard deviation would increase from 6% to 57.2%. Whether an investor should take the position offering the higher expected return would depend on the investor's preference for return relative to risk. We have more to say about this in Chapter 10.

In Figure 5.11 we have constructed the diagram for combinations of Colonel Motors and Separated Edison, assuming a correlation coefficient of 0.5. Notice that all portfolios offering returns above the global minimum variance portfolio lie along a concave curve. The reasoning for this is directly analogous to that presented when short sales were not allowed.

When we extend this analysis to the efficient frontiers of all securities and portfolios, we get a figure such as Figure 5.12, where *MVBC* is the efficient set. Because combinations of two portfolios are concave, the efficient set is concave. The efficient set still starts with the minimum variance portfolio, but when short sales are allowed, it has no finite upper bound.[13]

THE EFFICIENT FRONTIER WITH RISKLESS LENDING AND BORROWING

Up to this point we have been dealing with portfolios of risky assets. The introduction of a riskless asset into our portfolio possibility set considerably simplifies the analysis. We can consider lending at a riskless rate as investing in an asset with a certain outcome

[13]Merton (1972) has shown that the efficient set is the upper half of a hyperbola.

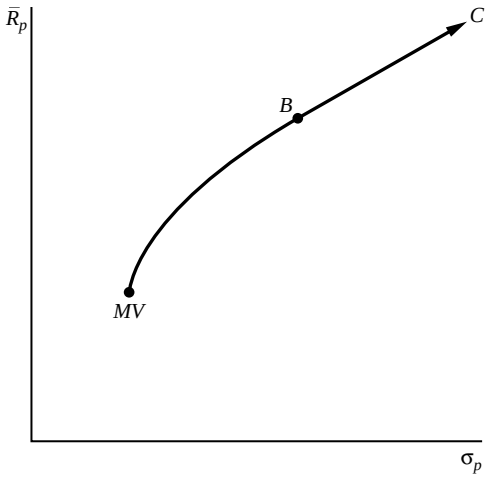


Figure 5.12 The efficient set when short sales are allowed.

(e.g., a short-term government bill or savings account). Borrowing can be considered as selling such a security short; thus borrowing can take place at the riskless rate.

We call the certain rate of return on the riskless asset R_F. Because the return is certain, the standard deviation of the return on the riskless asset must be zero.

We first examine the case where investors can lend and borrow unlimited amounts of funds at the riskless rate. Initially assume that the investor is interested in placing part of the funds in some portfolio *A* and either lending or borrowing. Under this assumption, we can easily determine the geometric pattern of all combinations of portfolio *A* and lending or borrowing. Call *X* the fraction of original funds that the investor places in portfolio *A*. Remember that *X* can be greater than 1 because we are assuming that the investor can borrow at the riskless rate and invest more than his initial funds in portfolio *A*. If *X* is the fraction of funds the investor places in portfolio *A*, $(1 - X)$ must be the fraction of funds that were placed in the riskless asset. The expected return on the combination of riskless asset and risky portfolio is given by

$$\bar{R}_C = (1-X)R_F + X\bar{R}_A$$

The risk on the combination is

$$\sigma_C = \left[(1-X)^2\sigma_F^2 + X^2\sigma_A^2 + 2X(1-X)\sigma_A\sigma_F\rho_{FA}\right]^{1/2}$$

Because we have already argued that σ_F is zero,

$$\sigma_C = \left(X^2\sigma_A^2\right)^{1/2} = X\sigma_A$$

Solving this expression for *X* yields

$$X = \frac{\sigma_C}{\sigma_A}$$

Substituting this expression for *X* into the expression for expected return on the combination yields

$$\bar{R}_C = \left(1 - \frac{\sigma_C}{\sigma_A}\right)R_F + \frac{\sigma_C}{\sigma_A}\bar{R}_A$$

Rearranging terms,

$$\bar{R}_C = R_F + \left(\frac{\bar{R}_A - R_F}{\sigma_A} \right) \sigma_C$$

Note that this is the equation of a straight line. All combinations of riskless lending or borrowing with portfolio *A* lie on a straight line in expected return standard deviation space. The intercept of the line (on the return axis) is R_F, and the slope is $(\bar{R}_A - R_F)/\sigma_A$. Furthermore, the line passes through the point $(\sigma_A, \bar{R}_A)$. This line is shown in Figure 5.13. Note that to the left of point *A*, we have combinations of lending and portfolio *A*, whereas to the right of point *A*, we have combinations of borrowing and portfolio *A*.

The portfolio *A* we selected for this analysis had no special properties. Combinations of any security or portfolio and riskless lending and borrowing lie along a straight line in expected return standard deviation of return space. Examine Figure 5.14. We could have combined portfolio *B* with riskless lending and borrowing and held combinations along the line R_FB rather than R_FA. Combinations along R_FB are superior to combinations along R_FA since they offer greater return for the same risk. It should be obvious that what we would like to do is to rotate the straight line passing through R_F as far as we can in a counterclockwise direction. The furthest we can rotate it is through point *G*.[14] Point *G* is the tangency point between the efficient frontier and a ray passing through the point R_F on the vertical axis. The investor cannot rotate the ray further because by the definition of the efficient frontier, there are no portfolios lying above the line passing through R_F and *G*.

All investors who believed they faced the efficient frontier and riskless lending and borrowing rates shown in Figure 5.14 would hold the same portfolio of risky assets—portfolio *G*. Some of these investors who were very risk averse would select a portfolio along the

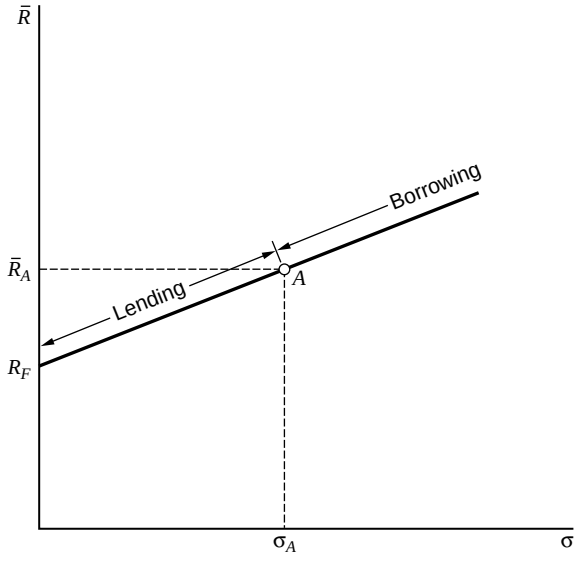


Figure 5.13 Expected return and risk when the risk-free rate is mixed with portfolio *A*.

[14]In this section we have drawn the efficient frontier as it would look if short sales were not allowed. However, the analysis is general and applies equally well to the case where short sales are allowed.

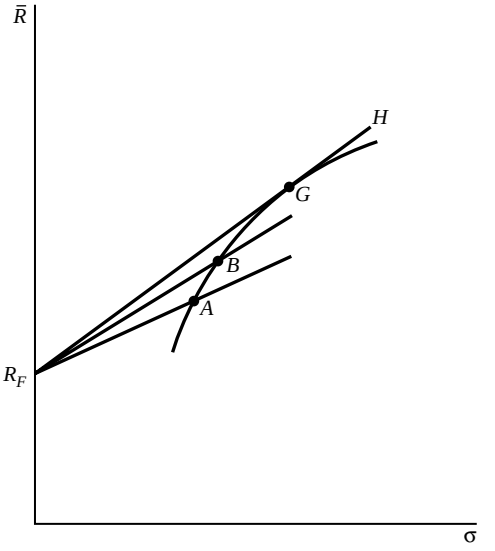


Figure 5.14 Combinations of the riskless asset and various risky portfolios.

segment R_F—G and place some of their money in a riskless asset and some in risky portfolio G. Others who were much more tolerant of risk would hold portfolios along the segment G—H, borrowing funds and placing their original capital plus the borrowed funds in portfolio G. Still other investors would just place the total of their original funds in risky portfolio G. All of these investors would hold risky portfolios with the exact composition of portfolio G. Thus, for the case of riskless lending and borrowing, identification of portfolio G constitutes a solution to the portfolio problem. The ability to determine the optimum portfolio of risky assets without having to know anything about the investor has a special name. It is called the *separation theorem*.[15]

Let us for a moment examine the shape of the efficient frontier under more restrictive assumptions about the ability of investors to lend and borrow at the risk-free rate. There is no question about the ability of investors to lend at the risk-free rate (buy government securities). If they can lend but not borrow at this rate, the efficient frontier becomes R_F—G—H in Figure 5.15. Certain investors will hold portfolios of risky assets located between G and H. However, any investor who held some riskless asset would place all remaining funds in the risky portfolio G.

Another possibility is that investors can lend at one rate but must pay a different and presumably higher rate to borrow. Calling the borrowing rate R'_F, the efficient frontier would become R_F—G—H—I in Figure 5.16. Here there is a small range of risky portfolios that would be optional for investors to hold. If R_F and R'_F are not too far apart, the assumption of riskless lending and borrowing at the same rate might provide a good approximation to the optimal range G—H of risky portfolios that investors might consider holding.

[15]The term *separation theorem* has, at times, been used to describe other phenomena in finance. We continue to use it in the preceding sense throughout this book.

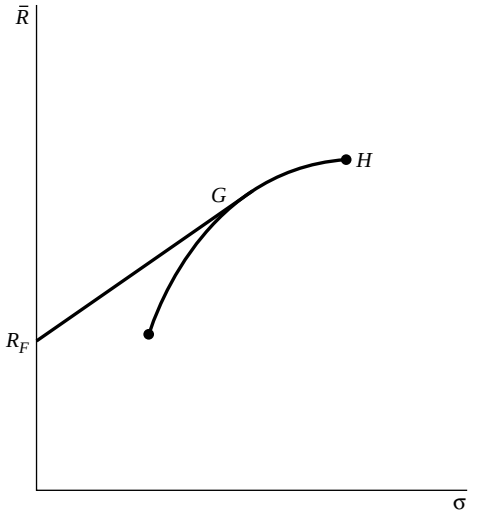


Figure 5.15 The efficient frontier with lending but not borrowing at the riskless rate.

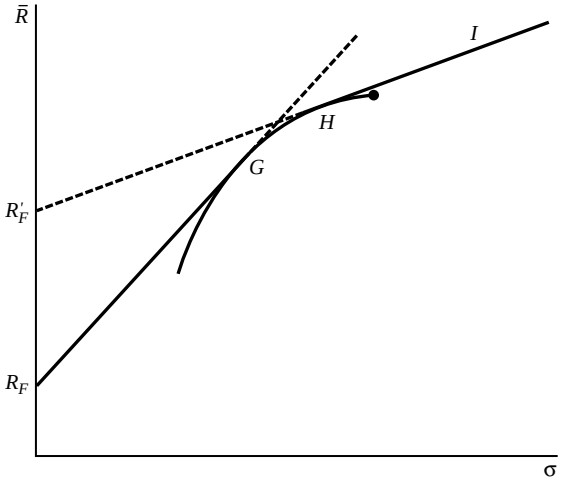


Figure 5.16 The efficient frontier with riskless lending and borrowing at different rates.

EXAMPLES AND APPLICATIONS

In this section we discuss some considerations that affect the choice of inputs to the portfolio selection problem and provide some examples of the use of the analysis just presented.

Considerations in Determining Inputs

Almost all asset allocation analysis starts out by estimating some of the inputs to the portfolio selection process using historical data. Analysts usually modify these historical estimates so that they better reflect beliefs about the future. In Chapters 7 and 8, on index models, we discuss ways of using historical data to obtain estimates of variances

and correlation coefficients that are more accurate than simply taking their historical value. However, before we do so, we discuss some general considerations in using historical data.

Inflation-Adjusted Inputs to Optimization The efficient frontier technology is widely used in practice to make asset allocation decisions for long-term investment, particularly for pension fund assets. When the investment horizon is measured in decades, it is important to consider how the change in the purchasing power of currency affects investment choice. In particular, investors may care more about the future purchasing-power value of the portfolio—that is, the value after adjusting for the effects of inflation—than about the future nominal value of the portfolio. One approach to this problem is to apply the efficient frontier technology to *inflation-adjusted* returns. Table 5.6 compares historical statistics for U.S. stocks, government bonds, Treasury bills, and inflation. Notice that Treasury bill returns are correlated with inflation and have a larger return when inflation is higher and a lower return when inflation is lower. This suggests that Treasury bills may serve as a partial inflation hedge.

Table 5.7 reports statistics for inflation-adjusted returns to stocks, bonds, and Treasury bills. The reader can use these inputs as a starting point when creating an efficient frontier for inflation-adjusted returns.

Although some securities, such as Treasury bills, provide a partial hedge against inflation, there is no "riskless" asset in the preceding example—even Treasury bills have some exposure to inflation. One security recently developed in the United States and used for some time in other countries, such as the United Kingdom, provides a near-perfect inflation hedge. Since 1997, the United States has issued inflation-linked securities whose value is determined, in part, by changes in the Consumer Price Index (an inflation measure). The return of these bonds varies with inflation, making the bonds a good hedge against inflation. At times, inflation has been over 10% per year in the United States, which means that wealth invested in assets uncorrelated to changes in inflation effectively

Table 5.6 Returns with No Inflation Adjustment

			Correlations			
	Arithmetic Mean	Standard Deviation	Stocks	Bonds	T-Bills	Inflation
Stocks	11.8	20.3	1.00			
Bonds	6.4	8.4	0.16	1.00		
T-Bills	3.6	3.1	−0.01	0.16	1.00	
Inflation	3.1	4.2	0.00	−0.16	0.41	1.00

Table 5.7 Returns after Adjusting for Inflation

	Arithmetic Mean	Standard Deviation	S&P	Bonds	T-Bills
S&P	8.6	20.3	1.00		
Bonds	3.4	9.5	0.22	1.00	
T-Bills	0.6	3.9	0.10	0.54	1.00

loses 10% of its purchasing power per year. Thus inflation-linked securities have the potential to protect against serious erosion of investor wealth in inflationary times.

In performing investment analysis the analyst may well want to examine inflation-adjusted returns along with or instead of nominal returns. Furthermore, inflation-linked securities are increasingly likely to be an important asset class in portfolio optimization.

Input Estimation Uncertainty Reliable inputs are crucial to the proper use of mean–variance optimization in the asset allocation decision. It is common to use historical risk, return, and correlation as a starting point in obtaining inputs for calculating the efficient frontier. If return characteristics do not change through time, then the longer the data are available, the more accurate is the estimate of the mean. To see this, note that the formula for the standard error of the mean of a sequence of independent random variables is $\frac{\sigma^2}{N}$, where N is the sample size. For a sequence of independent returns observed through time, N is the number of time periods since the beginning of the historically observed data. Thus, under the assumption of stationary (or unchanging) expected returns and returns uncorrelated through time, more historical data will improve the estimate of expected return included in the mean–variance model, although the improvement is diminishing.

To illustrate the importance of this issue for portfolio choice, imagine that the investor is forced to choose between two investments, each with identical sample means and variances. Other things equal, the standard approach would view the two investments as equivalent. If you consider the additional information that the first sample mean was based on 1 year of data and the second on 10 years of data, common sense would suggest that the second alternative is less risky than the first. Furthermore, we can assume that the investor is mainly concerned about next month's return, which has a mean return of $\bar{R}$ and a variance of $\sigma^2_{Pred} = \sigma^2 + \frac{\sigma^2}{T}$, where

σ^2_{Pred} is the predicted variance series
σ^2 is the variance of monthly return
T is the number of time periods

The first part of the expression captures the inherent risk in the return. The second term captures the uncertainty that comes from lack of knowledge about the true mean return. In a Bayesian analysis, the sum of the two terms on the right-hand side of this equation is referred to as the variance of the *predictive distribution* of returns. Notice that predicted variance is always greater than historical variance because of uncertainty as to the future mean.

Characteristics of security returns usually change over time. Thus there is a trade-off between using a long time frame to improve the estimates and having potentially inaccurate estimates from the longer time period because the security characteristics have changed. Because of this conflict, most analysts modify historical estimates to reflect their beliefs about how current conditions differ from past conditions.

The choice of the time period is more complicated when a relatively new asset class is added to the mix, and the available data for the new asset are much less than for other assets. For example, consider the addition of the International Financial Corporation's (IFC) index of emerging equity markets, which is available from 1985. An analyst who wishes to use historical data as a starting point for optimization could use all available data for calculating means, standard deviations, and correlations or data from only the common period of observation. Applying the first approach to U.S. capital market data would mean using the entire historical data from 1926 to the present from stocks and bonds. The second approach

Table 5.8 Returns over Different Decades

	1970s	1980s	1990s	2002–2011
Large stocks	17.2	19.4	15.9	16.6
Small stocks	30.8	22.5	20.2	23.7
Long-term corporate	8.7	14.1	6.9	12.2
Treasury bills	0.6	0.9	0.4	0.5

would use only data on U.S. markets from 1985 to the present. Table 5.8 shows returns over different decades.

Notice that small stocks had their highest return in the 1970s, while large stocks had their highest returns in the 1980s. Furthermore, all asset classes had low returns in the 1990s. Thus which period we use to measure historical returns strongly affects our estimates and the optimum portfolio.

Does the correlation between stock and bond returns follow a predictable pattern that could help with input estimation? Li (2002) showed that the stock–bond correlation followed similar time trends across many countries. It reached a peak in 1996 of around 0.5 in most of the major industrialized countries except Japan. By 2002, the stock–bond correlation had turned negative. Why? Li found that this critical correlation changed with shifts in uncertainty about future inflation. As inflation uncertainty rises, the stock–bond correlation rises as well. The correlation among international equity markets changes significantly through time also. The average correlation between major stock markets over the past 150 years has ranged from less than 10% (1880s and 1890s, and 1940–1980) to over 30% (1860s, 1930s, 1990s). Goetzmann, Li, and Rouwenhorst (2005) studied the relationship between globalization and market correlations over this time period. They attribute the higher correlations among equity markets to periods of greater liberalization in cross-border flows. The result of research on time variation in correlations suggests that macroeconomic conditions may have an effect on correlation forecasts, which indeed appears to be the case (Brown et al., 2009).

Short-Horizon Inputs and Long-Horizon Portfolio Choice Another important consideration in estimating inputs to the optimization process is the effect of the investment time horizon on variance. In the previous example we saw that under the assumption that returns were uncorrelated from one period to the next, the standard error of the mean decreased with the square root of time. This is based on a more general result that the sum of the variance of a sequence of random variables is equal to the variance of the sum. When actual returns are examined, some securities have returns that are highly correlated over time (e.g., autocorrelated). Treasury bill returns, for example, tend to be highly autocorrelated, meaning that the return to investing in T-bills in one year does a good job at predicting the return to investing in T-bills the next year. High T-bill returns are more likely to be followed by high returns than low returns. Thus, although the standard deviation of T-bills is low over short intervals, on a percentage basis, it significantly increases as the time period of observation increases from 1 to 5 to 10 years. Thus T-bills are effectively an increasingly risky asset as the investment time horizon grows. For example, research by Edwards and Goetzmann (1994) shows that the estimated annualized standard deviation for Treasury bill returns over the 10-year horizon is about 6%, compared to the 3.2% annual standard deviation measured at the 1-year horizon.

THREE EXAMPLES

Let us return to the two examples discussed in Chapter 4. Consider first the allocation between equity and debt. The minimum variance portfolio is given by Equation (5.9). The estimated inputs for bonds and stocks are

$$\bar{R}_S = 11.8\% \quad \sigma_S = 20.3\% \quad \rho_{SB} = 0.01$$
$$\bar{R}_B = 6.4\% \quad \sigma_S = 8.4\%$$

Plugging the values for standard deviation and correlation into Equation (5.9) gives

$$X_S = \frac{(8.4)^2 - 0.01(8.4)(20.3)}{(8.4)^2 + (20.3)^2 - 2(0.01)(8.4)(20.3)}$$
$$X_S = 0.144 = 14.4\%$$

Thus the minimum variance portfolio involves 14.4% stock. The associated standard deviation is 7.8%, which is slightly less than the standard deviation associated with investing 100% in bonds. The dots in Figure 5.17 are plots of all combinations of a stock index and bond index. As we move to the right, each dot represents the expected return and standard deviation of a portfolio with 10% more in common stock. This is the efficient frontier with short sales allowed (although it would continue to the right). The efficient frontier with no short sales is shown in Figure 4.4.

At the time of this revision the interest rate on Treasury bills was about 1%. Using this as a riskless lending and borrowing rate, the tangency portfolio is portfolio *T* shown in Figure 5.17. We will see how this is calculated in the next chapter. The expected return and risk for portfolio *T* as read from the graph is 7.75% and 8.13%, respectively. Thus the slope of the line connecting the tangency portfolio and the efficient frontier is

$$\frac{7.75 - 1}{8.13} = 0.83$$

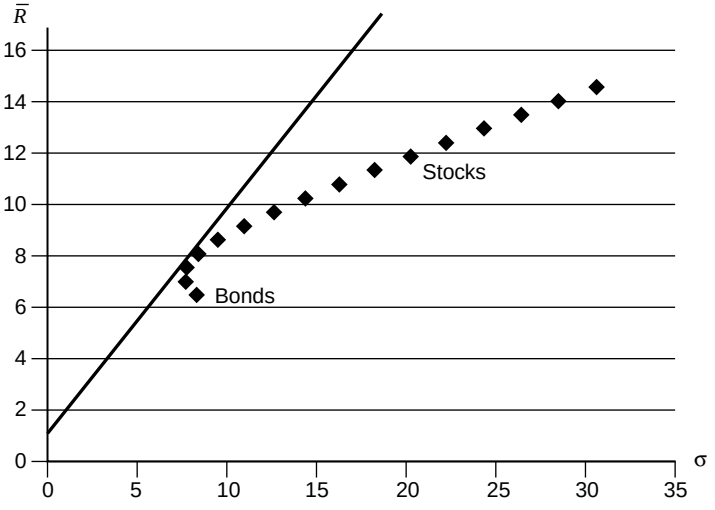


Figure 5.17 The efficient frontier of stocks and bonds.

and the equation of the efficient frontier with riskless lending and borrowing is

$$\overline{R}_P = 1 + 0.83p$$

Once we know the expected return of portfolio *T*, we can easily determine its composition. Simply recall that

$$\overline{R}_P = X_S \overline{R}_S + (1 - X_S)\overline{R}_B$$

Therefore

$$7.75 = X_S(11.8) + (1 - X_S)6.4$$

and

$$X_{S\&P} = 0.25 \qquad X_B = 0.75$$

The second example we examined in Chapter 4 was a combination of a domestic stock portfolio and an international portfolio. All combinations of these two funds without short sales were represented by Figure 4.5. Note that part of these combinations is inefficient. The estimated inputs were

$$\overline{R}_S = 11.8\% \quad \sigma_S = 20.3\% \quad \rho_{S\,\text{int}} = 0.66$$
$$\overline{R}_{\text{int}} = 9.2\% \quad \sigma_{\text{int}} = 18.4\%$$

Solving for the global minimum variance portfolio, we have

$$X_S = \frac{(18.4)^2 - 0.66\,(18.4)\,(20.3)}{(18.4)^2 + (20.3)^2 + 2(0.66)(18.4)(20.3)}$$
$$X_S = 36$$

Thus the global minimum variance portfolio is obtained by investing 36% in the domestic index and 64% in the foreign portfolio. The resulting standard deviation is 17.5%, which is less than the standard deviation of both portfolios. This is an example of how diversification can reduce risk. Note that it is inefficient to hold the foreign portfolio by itself. An investor wishing to accept the risk of 18.4% on the foreign portfolio could obtain an expected return of 10.2% by putting 37% in the stock index and 63% in the foreign portfolio. Thus, at an 18.4% standard deviation, the increase in expected return from using the optimum combination is 1% with no increase in risk. The efficient frontier with no short sales is the scatter of dots in Figure 5.18 from 100% in bonds to 150% in the domestic stock index. Each dot as we move to the right represents the expected return and standard deviation of return as we increase the amount of the domestic stock index by 10%. The efficient frontier with short sales allowed is the complete scatter of dots shown in Figure 5.18 (although it would continue to the right).

If the riskless lending and borrowing rate is 1%, then the tangency portfolio is 70% in the domestic portfolio and 30% in the international portfolio. The associated mean return is 11%, and standard deviation of return is 18.3%. Thus the slope of the efficient frontier with riskless lending and borrowing is

$$\frac{11 - 1}{18.3} = 0.552$$

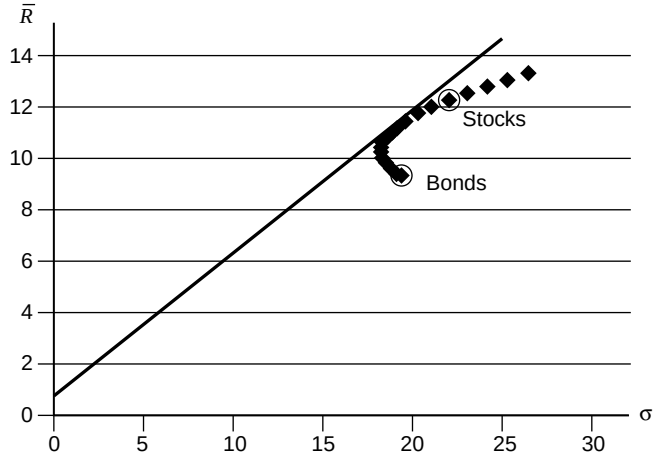


Figure 5.18 The efficient frontier of domestic and international stocks.

The equation of the efficient frontier is

$$\overline{R}_P = 1 + 0.552\sigma_P$$

As a third example, consider the asset allocation problem across bonds, domestic stocks, and international stocks. We continue to use all the inputs from the prior examples. We need one additional input, the correlation coefficient between bonds and the international portfolio. Past data indicate a value of 0.05 is reasonable. Various combinations of these three assets, some of which lie on the efficient frontier and some of which do not, are plotted as dots in Figure 5.19. Note that both the international portfolio and the bond portfolio are obviously dominated by other portfolios. The figure does not include portfolios involving short sales. Thus, because the stock index has the highest expected return, it is not dominated. The efficient frontier would be the dots that have the highest mean return for a given standard deviation.

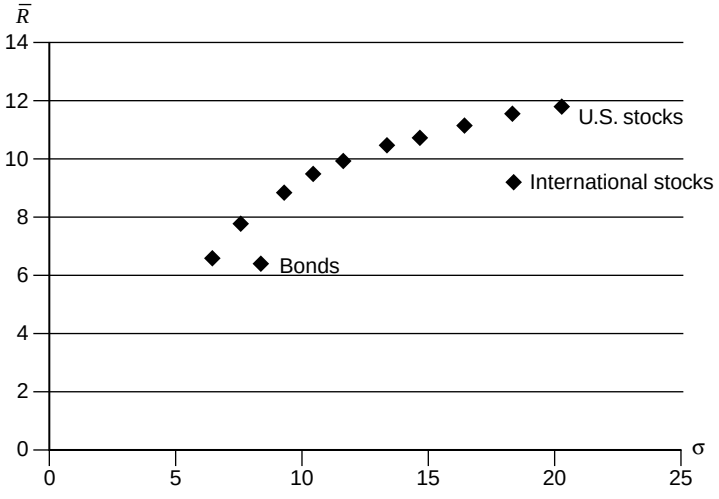


Figure 5.19 Combinations of bonds, U.S. stocks, and international stocks.

The tangency portfolio with a riskless lending and borrowing rate has the following proportions:[16]

$$X_{SP} = 0.21$$
$$X_B = 0.73$$
$$X_{\text{int}} = 0.06$$

The expected return of this portfolio is 9.5%, and the standard deviation is 10.5%. Thus the slope of the efficient frontier with riskless lending and borrowing is 0.905 and the equation of the efficient frontier is

$$\overline{R}_P = 1 + 0.905\sigma_P$$

Compare this to the efficient frontier derived with two risky assets and a riskless asset. This efficient frontier dominates the efficient frontier using only the S&P and bonds as the risky assets as well as the efficient frontier using only the S&P and international stocks as risky assets.

CONCLUSION

In this chapter we have defined the geometric properties of that set of portfolios all risk-avoiding investors would hold regardless of their specific tolerance for risk. We have defined this set—the efficient frontier—under alternative assumptions about short sales and the ability of the investor to lend and borrow at the riskless rate. Now that we understand the geometric properties of the efficient frontier, we are in a position to discuss solution techniques to the portfolio problem. This is done in the following chapter.

QUESTIONS AND PROBLEMS

1. Return to the example presented in Problem 1, Chapter 4.

A. Assuming short selling is not allowed:

(1) For securities 1 and 2, find the composition, standard deviation, and expected return of the portfolio that has minimum risk.

(2) On the same graph, plot the expected return and standard deviation for all possible combinations of securities 1 and 2.

(3) Assuming that investors prefer more to less and are risk avoiders, indicate in red those sections of the diagram in Part 2 that are efficient.

(4) Repeat steps 1, 2, and 3 for all other possible pairwise combinations of the securities shown in Problem 1 of Chapter 4.

B. Assuming short selling is allowed:

(1) For securities 1 and 2, find the composition, standard deviation, and expected return of the portfolio that has minimum risk.

(2) On the same graph, plot the expected return and standard deviation for all possible combinations of securities 1 and 2.

(3) Assuming that investors prefer more to less and are risk avoiders, indicate in red those sections of the diagram in Part 2 that are efficient.

[16]Techniques for obtaining this solution are presented in Chapter 6.

(4) Repeat steps 1, 2, and 3 for all other possible pairwise combinations of the securities shown in Problem 1 of Chapter 4.

C. Assuming that the riskless lending and borrowing rate is 5%, and short sales are allowed, find the location of the optimal portfolio from among those considered. Repeat for a rate of 8%.

2. Answer the questions to Problem 1 with data from Chapter 4, Problem 2.

3. For Problem 2, find the composition of the portfolio that has minimum variance for each of the two security combinations you considered.

4. Derive the expression for the location of all portfolios of two securities in expected return standard deviation space when the correlation between the two securities is -1.

5.

	Expected Return	Standard Deviation
Security 1	10%	5%
Security 2	4%	2%

For the two securities shown, plot all combinations of the two securities in $\bar{R}_p$ σ_p space. Assume $\rho = 1, -1, 0$. For each correlation coefficient, what is the combination that yields the minimum σ_p and what is that σ_p? Assume no short selling.

6. In Problem 5, assume a riskless rate of 10%. What is the optimal investment?

BIBLIOGRAPHY

1. Bawa, Vijay. "Admissible Portfolios for All Individuals," *Journal of Finance*, **XXXI,** No. 3 (Sept. 1976), pp. 1169–1183.
2. Ben-Horim, Moshe, and Levy, Haim. "Total Risk, Diversifiable Risk and Non-Diversifiable Risk: A Pedagogic Note," *Journal of Financial and Quantitative Analysis*, **XV,** No. 2 (June 1980), pp. 289–298.
3. Brennan, Michael J., and Kraus, Allan. "The Geometry of Separation and Myopia," *Journal of Financial and Quantitative Analysis*, **XI,** No. 2 (June 1976), pp. 171–193.
4. Brown, Stephen, and Barry, Christopher. "Differential Information and the Small Firm Effect," *Journal of Financial Economics*, **13** (1984), pp. 283–294.
5. ——. "Differential Information and Security Market Equilibrium," *Journal of Financial and Quantitative Analysis*, **20** (1985), pp. 407–422.
6. Brown, Stephen, Hiraki, Takato, Arakawa, Kiyoshi, and Ohno, Saburo. "Risk Premia in International Equity Markets Revisited," *Pacific-Basin Finance Journal* **17,** No. 3 (2009) pp. 295–318 (forthcoming).
7. Brumelle, Shelby. "When Does Diversification between Two Investments Pay?" *Journal of Financial and Quantitative Analysis*, **IX,** No. 3 (June 1974), pp. 473–483.
8. Buser, Stephen. "A Simplified Expression for the Efficient Frontier in Mean–Variance Portfolio Analysis," *Management Science*, **23** (April 1977), pp. 901–903.
9. Canner, Niko. "An Asset Allocation Puzzle," *American Economic Review*, **87,** No. 1 (March 1997), pp. 181–193.
10. Cass, David, and Stiglitz, Joseph. "The Structure of Investor Preferences and Asset Returns, and Separability in Portfolio Allocation: A Contribution to the Pure Theory of Mutual Funds," *Journal of Economic Theory*, **2,** No. 2 (June 1970), pp. 122–160.
11. Dalal, Ardeshir J. "On the Use of a Covariance Function in a Portfolio Model," *Journal of Financial and Quantitative Analysis*, **XVIII,** No. 2 (June 1983), pp. 223–228.
12. Edwards, Franklin, and Goetzmann, William. "Short Horizon Inputs and Long Horizon Portfolio Choice," *Journal of Portfolio Management*, **20,** No. 4 (Summer 1994), pp. 76–81.

13. Elton, Edwin J., and Gruber, Martin J. "Dynamic Programming Applications in Finance," *Journal of Finance*, **XXVI,** No. 2 (May 1971), pp. 473–505.
14. ——."Portfolio Theory When Investment Relatives Are Lognormally Distributed," *Journal of Finance*, **XXIX,** No. 4 (Sept. 1974), pp. 1265–1273.
15. Friedman, Harris. "Real Estate Investment and Portfolio Theory," *Journal of Financial and Quantitative Analysis*, **VI,** No. 2 (March 1971), pp. 861–873.
16. Gibbons, Michael R., and Shanken, Jay. "Subperiod Aggregation and the Power of Multivariate Tests of Portfolio Efficiency," *Journal of Financial Economics*, **19,** No. 2 (Dec. 1987), pp. 389–394.
17. Goetzmann, William, Li, Lingfeng, and Rouwenhorst, K. Geert. "Long-Term Global Market, Correlations," *The Journal of Business*, **78**, No. 1 (2005), pp. 1–38.
18. Grauer, Robert R., and Hakansson, Nils H. "A Half Century of Returns on Levered and Unlevered Portfolios of Stocks, Bonds, and Bills, with and without Small Stocks," *Journal of Business*, **59,** No. 2 (April 1986), p. 287.
19. Hakansson, Nils. "Risk Disposition and the Separation Property in Portfolio Selection," *Journal of Financial and Quantitative Analysis*, **IV,** No. 4 (Dec. 1969), pp. 401–416.
20. ——. "An Induced Theory of the Firm under Risk: The Pure Mutual Fund," *Journal of Financial and Quantitative Analysis*, **V,** No. 2 (May 1970), pp. 155–178.
21. Li, Lingfeng. "Macroeconomic Factors and the Correlation of Stock and Bond Returns," Yale ICF working paper No. 02-46 (Nov. 2002).
22. Merton, Robert. "An Analytic Derivation of the Efficient Portfolio Frontier," *Journal of Financial and Quantitative Analysis*, **VII,** No. 4 (Sept. 1972), pp. 1851–1872.
23. Mossin, Jan. "Optimal Multiperiod Portfolio Policies," *Journal of Business*, **41,** No. 2 (April 1968), pp. 215–229.
24. Ohlson, James. "Portfolio Selection in a Log-Stable Market," *Journal of Financial and Quantitative Analysis*, **X,** No. 2 (June 1975), pp. 285–298.
25. Ohlson, J. S., and Ziemba, W. T. "Portfolio Selection in a Lognormal Market When the Investor Has a Power Utility Function," *Journal of Financial and Quantitative Analysis*, **XI,** No. 1 (March 1976), pp. 57–71.
26. Pye, Gordon. "Lifetime Portfolio Selection in Continuous Time for a Multiplicative Class of Utility Functions," *American Economic Review*, **LXIII,** No. 5 (Dec. 1973), pp. 1013–1020.
27. Rosenberg, Barr, and Ohlson, James. "The Stationarity Distribution of Returns and Portfolio Separation in Capital Markets: A Fundamental Contradiction," *Journal of Financial and Quantitative Analysis*, **XI,** No. 3 (June 1973), pp. 393–401.
28. Shanken, Jay. "A Bayesian Approach to Testing Portfolio Efficiency," *Journal of Financial Economics*, **19,** No. 2 (Dec. 1987), pp. 195–216.
29. Smith, Keith. "Alternative Procedures for Revising Investment Portfolios," *Journal of Financial and Quantitative Analysis*, **III,** No. 4 (Dec. 1968), pp. 371–403.
30. Zhou, Guofu. "Small Sample Tests of Portfolio Efficiency," *Journal of Financial Economics*, **30,** No. 1 (Nov. 1991), pp. 165–192.

6

Techniques for Calculating the Efficient Frontier

In Chapters 4 and 5 we discussed the properties of the efficient frontier under alternative assumptions about lending and borrowing and alternative assumptions about short sales. In this chapter we describe and illustrate methods that can be used to calculate efficient portfolios. By necessity, this chapter is more mathematically complex than those that preceded it and most of those that follow. The reader who is concerned only with a conceptual approach to portfolio management can skip this chapter and still understand later ones. However, we believe that knowledge of the solution techniques to portfolio problems outlined here yields a better understanding and appreciation of portfolio management.

We have not followed the same order in presenting solution techniques for portfolio problems as was followed in describing the properties of the efficient set (Chapter 5). Rather, we have rearranged the order so that solution techniques are presented from the simplest to the most complex. The first four sections of this chapter discuss the solution to the portfolio problem when it is assumed in turn that

1. short sales are allowed and riskless lending and borrowing is possible

2. short sales are allowed but riskless lending or borrowing is not permitted

3. short sales are disallowed but riskless lending and borrowing exists

4. neither short sales nor riskless lending and borrowing is allowed

A fifth section shows how additional constraints, such as the need for a minimum dividend yield, can be incorporated into the portfolio problem. The solution techniques discussed here are the ones used in actual applications. For most problems, the calculations are lengthy enough that computers are used. Indeed, computer programs exist for each of the techniques discussed. In addition, in Chapter 9, we present simplifications of the procedures discussed in the present chapter that are useful in solving most real problems. This chapter is necessary for an understanding of the computer programs and an appreciation of the simple rules discussed later. Thus, although this chapter is more demanding than some others, it is well worth the effort needed to understand it.

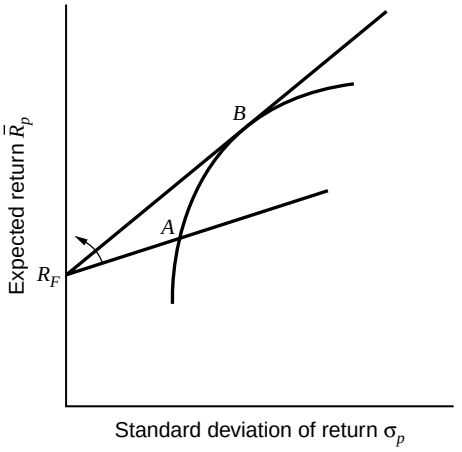


Figure 6.1 Combinations of the riskless asset in a risky portfolio.

SHORT SALES ALLOWED WITH RISKLESS LENDING AND BORROWING

The derivation of the efficient set when short sales are allowed and there is a riskless lending and borrowing rate is the simplest case we can consider. From Chapter 5 we know that the existence of a riskless lending and borrowing rate implies that there is a single portfolio of risky assets that is preferred to all other portfolios. Furthermore, in return standard deviation space, this portfolio plots on the ray connecting the riskless asset and a risky portfolio that lies furthest in the counterclockwise direction. For example, in Figure 6.1, the portfolio on the ray R_F—B is preferred to all other portfolios of risky assets. The efficient frontier is the entire length of the ray extending through R_F and B. Different points along the ray R_F—B represent different amounts of borrowing and/or lending in combination with the optimum portfolio of risky assets B.

An equivalent way of identifying the ray R_F—B is to recognize that it is the ray with the greatest slope. Recall that the slope of the line connecting a riskless asset and a risky portfolio is the expected return on the portfolio minus the risk-free rate divided by the standard deviation of the return on the portfolio. Thus the efficient set is determined by finding that portfolio with the greatest ratio of excess return (expected return minus risk-free rate) to standard deviation that satisfies the constraint that the sum of the proportions invested in the assets equals 1. In equation form we have the following: maximize the objective function

$$\theta = \frac{\bar{R}_P - R_F}{\sigma_P}$$

subject to the constraint[1]

$$\sum_{i=1}^{N} X_i = 1$$

[1]Lintner (1965) has advocated an alternative definition of short sales, one that is more realistic. He assumes correctly that when an investor sells stock short, cash is not received but rather is held as collateral. Furthermore, the investor must put up an additional amount of cash equal to the amount of stock he or she sells short. The investor generally does not receive any compensation (interest) on these funds. However, if the investor is a broker-dealer, interest can be earned on both the money put up and the money received from the short sale of securities. As shown in Appendix A, this leads to the constraint $\Sigma|X_i| = 1$ and leaves all other equations unchanged.

This is a constrained maximization problem. There are standard solution techniques available for solving it. For example, it can be solved by the method of Lagrangian multipliers. There is an alternative. The constraint could be substituted into the objective function and the objective function maximized as in an unconstrained problem. This latter procedure will be followed subsequently. We can write R_F as R_F times 1. Thus we have

$$R_F = 1R_F = \left(\sum_{i=1}^{N} X_i\right) R_F = \sum_{i=1}^{N} \left(X_i R_F\right)$$

Making this substitution in the objective function and stating the expected return and standard deviation of return in the general form, derived in Chapter 4, yields

$$\theta = \frac{\sum_{i=1}^{N} X_i\left(\overline{R}_i - R_F\right)}{\left[\sum_{i=1}^{N} X_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \sigma_{ij}\right]^{1/2}}$$

The problem stated previously is a very simple maximization problem and as such can be solved using the standard methods of basic calculus. In calculus it is shown that to find the maximum of a function, you take the derivative with respect to each variable and set it equal to zero.[2] Thus the solution to the maximization problem just presented involves finding the solution to the following system of simultaneous equations:

$$\begin{aligned}
&1. \quad \frac{d\theta}{dX_1} = 0 \\
&2. \quad \frac{d\theta}{dX_2} = 0 \\
&3. \quad \frac{d\theta}{dX_3} = 0 \\
&\vdots \\
&N. \quad \frac{d\theta}{dX_N} = 0
\end{aligned}$$

In Appendix B at the end of this chapter we show that

$$\begin{aligned}
\frac{d\theta}{dX_i} = -\big(&\lambda X_1 \sigma_{1i} + \lambda X_2 \sigma_{2i} + \lambda X_3 \sigma_{3i} + \cdots + \lambda X_i \sigma_i^2 + \cdots \\
&+ \lambda X_{N-1} \sigma_{N-1i} + \lambda X_N \sigma_{Ni}\big) + \overline{R}_i - R_F = 0
\end{aligned}$$

[2]Solving the problem without constraining the solution by

$$\sum_{i=1}^{N} X_i = 1$$

does not work in every maximization problem. It works here because the equations are homogeneous of degree zero.

where λ is a constant.[3] A mathematical trick allows a useful modification of the derivative. Note that each X_k is multiplied by a constant λ. Define a new variable $Z_k = \lambda X_k$. The X_k are the fraction to invest in each security, and the Z_k are proportional to this fraction. Substituting Z_k for the λX_k simplifies the formulation. To solve for the X_k after obtaining the Z_k, one divides each Z_k by the sum of the Z_k. Substituting Z_k for $\lambda_k X_k$ and moving the variance covariance terms to the right-hand side of the equality yields

$$\bar{R}_i - R_F = Z_1\sigma_{1i} + Z_2\sigma_{2i} + \cdots + Z_i\sigma_i^2 + \cdots + Z_{N-1}\sigma_{N-1i} + Z_N\sigma_N$$

We have one equation like this for each value of *i*. Thus the solution involves solving the following system of simultaneous equations:

$$\begin{aligned}
\bar{R}_1 - R_F &= Z_1\sigma_1^2 + Z_2\sigma_{12} + Z_3\sigma_{13} + \cdots + Z_N\sigma_{1N} \\
\bar{R}_2 - R_F &= Z_1\sigma_{12} + Z_2\sigma_2^2 + Z_3\sigma_{23} + \cdots + Z_N\sigma_{2N} \\
\bar{R}_3 - R_F &= Z_1\sigma_{13} + Z_2\sigma_{23} + Z_3\sigma_3^2 + \cdots + Z_N\sigma_{3N} \\
&\vdots \\
\bar{R}_N - R_F &= Z_1\sigma_{1N} + Z_2\sigma_{2N} + Z_3\sigma_{3N} + \cdots + Z_N\sigma_N^2
\end{aligned} \tag{6.1}$$

The *Z*s are proportional to the optimum amount to invest in each security. To determine the optimum amount to invest, we first solve the equations for the *Z*s. Note that this does not present a problem. There are *N* equations (one for each security) and *N* unknowns (the Z_k for each security). Then the optimum proportion to invest in stock *k* is X_k, where

$$X_k = Z_k \Big/ \sum_{i=1}^{N} Z_i$$

Let us solve an example. Consider three securities: Colonel Motors with expected return of 14% and standard deviation of return of 6%, Separated Edison with average return of 8% and standard deviation of return of 3%, and Unique Oil with mean return of 20% and standard deviation of return of 15%. Furthermore, assume that the correlation coefficient between Colonel Motors and Separated Edison is 0.5, between Colonel Motors and Unique Oil is 0.2, and between Separated Edison and Unique Oil is 0.4. Finally, assume that the riskless lending and borrowing rate is 5%. Equation (6.1) for three securities is

$$\begin{aligned}
\bar{R}_1 - R_F &= Z_1\sigma_1^2 + Z_2\sigma_{12} + Z_3\sigma_{13} \\
\bar{R}_2 - R_F &= Z_1\sigma_{12} + Z_2\sigma_2^2 + Z_3\sigma_{23} \\
\bar{R}_3 - R_F &= Z_1\sigma_{13} + Z_2\sigma_{23} + Z_3\sigma_3^2
\end{aligned}$$

Substituting in the assumed values, we get the following system of simultaneous equations:

$$\begin{aligned}
14 - 5 &= 36Z_1 + (0.5)(6)(3)Z_2 + (0.2)(6)(15)Z_3 \\
8 - 5 &= (0.5)(6)(3)Z_1 + 9Z_2 + (0.4)(3)(15)Z_3 \\
20 - 5 &= (0.2)(6)(15)Z_1 + (0.4)(3)(15)Z_2 + 225Z_3
\end{aligned}$$

[3]The constant is equal to $(\bar{R}_P - R_F)$ divided by σ_P^2.

Simplifying,

$$9 = 36Z_1 + 9Z_2 + 18Z_3$$
$$3 = 9Z_1 + 9Z_2 + 18Z_3$$
$$15 = 18Z_1 + 18Z_2 + 225Z_3$$

Further simplifying,

$$1 = 4Z_1 + Z_2 + 2Z_3$$
$$1 = 3Z_1 + 3Z_2 + 6Z_3$$
$$5 = 6Z_1 + 6Z_2 + 75Z_3$$

The solution to this system of simultaneous equations is

$$Z_1 = \frac{14}{63}, \qquad Z_2 = \frac{1}{63}, \qquad \text{and} \qquad Z_3 = \frac{3}{63}$$

The reader can verify this solution by substituting these values of Z_k into the foregoing equations.[4] The proportion to invest in each security is easy to determine. We know that each Z_k is proportional to X_k. Consequently, all we have to do to determine X_k is to scale the Z_k so that they add to 1.[5] For the foregoing problem,

$$\sum_{i=1}^{3} Z_i = \frac{18}{63}$$

Thus the proportion to invest in each security is

$$X_1 = \frac{14}{18}, \qquad X_2 = \frac{1}{18}, \qquad \text{and} \qquad X_3 = \frac{3}{18}$$

The expected return on the portfolio is

$$\overline{R}_P = \frac{14}{18}(14) + \frac{1}{18}(8) + \frac{3}{18}(20) = 14\tfrac{2}{3}\%$$

The variance of the return on the portfolio is[6]

$$\sigma_P^2 = \left(\frac{14}{18}\right)^2 (36) + \left(\frac{1}{18}\right)^2 9 + \left(\frac{3}{18}\right)^2 (225) + 2\left(\frac{14}{18}\right)\left(\frac{1}{18}\right)(6)(3)(0.5)$$
$$+ 2\left(\frac{14}{18}\right)\left(\frac{3}{18}\right)(6)(15)(0.2) + 2\left(\frac{1}{18}\right)\left(\frac{3}{18}\right)(3)(15)(0.4) = \frac{203}{6} = 33\tfrac{5}{6}$$

---

[4]See Appendix C at the end of this chapter for a description of solution techniques for systems of simultaneous equations.

[5]In the case of Lintnerian short sales, simply scale so that

$$\sum_{i=1}^{3} |X_i| = 1$$

[6]The variance of the portfolio could have been determined in another way. Recall that λ is the ratio of the excess return on the optimum portfolio divided by the variance of the optimum portfolio. Thus

$$\lambda = \frac{\overline{R}_P - R_F}{\sigma_P^2} = \frac{14\frac{2}{3} - 5}{\sigma_P^2}$$

Also recall that $Z_i = \lambda X_i$ so that $\Sigma Z_i = \lambda \Sigma X_i = \lambda$. Earlier we determined that $\Sigma Z_i = \lambda = 18/63$. Equating these two equations and solving for σ_P^2 yields the value presented earlier.

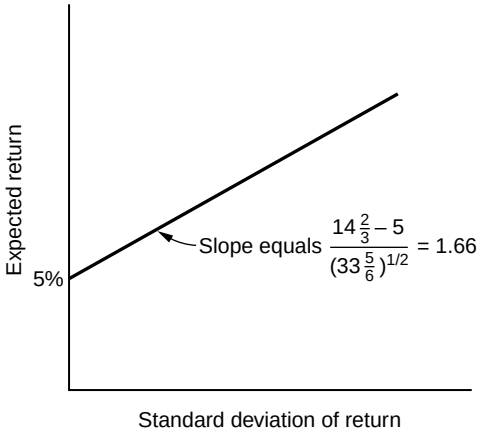


Figure 6.2 The efficient set with riskless lending and borrowing.

The efficient set is a straight line with an intercept at the risk-free rate of 5% and a slope equal to the ratio of excess return to standard deviation (see Figure 6.2). There are standard computer packages for the solution of a system of simultaneous equations. Appendix C at the end of this chapter presents some methods of solving them when the number of securities involved is limited so that hand calculations are reasonable.

SHORT SALES ALLOWED: NO RISKLESS LENDING AND BORROWING

When the investor does not wish to make the assumption that she can borrow and lend at the riskless rate of interest, the solution developed in the last section must be modified. However, much of the analysis can still be utilized. Consider Figure 6.3. The riskless lending and borrowing rate of 5% led to the selection of portfolio *B.* If the riskless lending and borrowing rate had been 4%, the investor would invest in portfolio *A.* If the investor's lending and borrowing rate was 6%, the investor would select portfolio *C.* These observations suggest the following procedure. Assume that a riskless lending and borrowing rate exists and find the optimum portfolio. Then assume that a different riskless lending and borrowing rate exists and find the optimum portfolio that corresponds to this second rate. Continue changing the assumed riskless rate until the full efficient frontier is determined.[7]

In Appendix D we present a general solution to this problem. We show that the optimal proportion to invest in any security is simply a linear function of R_F. Furthermore, because the entire efficient frontier can be constructed as a combination of any two portfolios that lie along it, the identification of the characteristics of the optimal portfolio for any two arbitrary values of R_F is sufficient to trace out the total efficient frontier.

RISKLESS LENDING AND BORROWING WITH SHORT SALES NOT ALLOWED

This problem is analogous to the case of riskless lending and borrowing with short sales allowed. One portfolio is optimal. Once again, it is the one that maximizes the slope of the line connecting the riskless asset and a risky portfolio. However, the set of portfolios that

[7]This works only for the standard definition of short sales. The Lintner definition of short sales assumes riskless lending and borrowing at a particular rate for each point on the original (curved) efficient frontier.

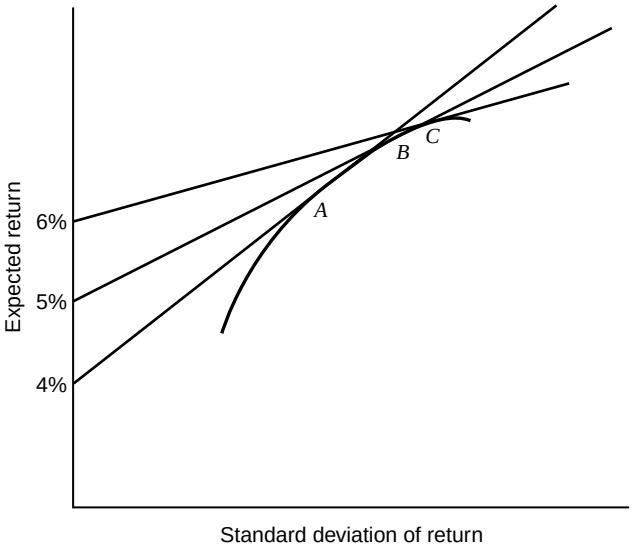


Figure 6.3 Tangency portfolios for different riskless rates.

is available to combine with lending and borrowing is different because a new constraint has been added. Investors cannot hold securities in negative amounts. More formally, the problem can be stated as

$$\text{Maximize } \theta = \frac{\overline{R}_P - R_F}{\sigma_P}$$

subject to

$$(1) \quad \sum_{i=1}^{N} X_i = 1$$

$$(2) \quad X_i \geq 0 \qquad \text{all } i$$

This is a mathematical programming problem because of the inequality restriction on X_i. At first glance, this might look like a linear programming problem. In fact, the constraints (1) and (2) are linear constraints. The problem is that the objective function (the expression we are maximizing) is not linear; σ_P contains terms involving X_i^2 and X_iX_j. Equations involving squared terms and cross-product terms are called quadratic equations. Since this looks like a linear programming problem, except that the objective function is quadratic rather than linear, it is called a *quadratic programming problem*. There are standard computer packages for solving quadratic programming problems, just as there are for linear programming problems, and the reader interested in solving a large-scale problem would utilize one of them. Some discussion of solution techniques is contained in Appendix E at the end of this chapter.

NO SHORT SELLING AND NO RISKLESS LENDING AND BORROWING

Recall that an efficient set is determined by minimizing the risk for any level of expected return. If we specify the return at some level and minimize risk, we have one point on the

efficient frontier. Thus, to get one point on the efficient frontier, we minimize risk subject to the return being some level plus the restriction that the sum of the proportions invested in each security is 1 and that all securities have positive or zero investment. This yields the following problem:

$$\text{Minimize} \sum_{i=1}^{N}\left(X_i^2\sigma_i^2\right)+\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N}\left(X_iX_j\sigma_{ij}\right)$$

subject to

$$(1)\quad \sum_{i=1}^{N} X_i = 1$$

$$(2)\quad \sum_{i=1}^{N}\left(X_i\bar{R}_i\right) = \bar{R}_P$$

$$(3)\quad X_i \geq 0, \qquad i = 1,\ldots,N$$

Varying $\bar{R}_P$ between the return on the minimum variance portfolio and the return on the maximum return portfolio traces out the efficient set. Once again, the problem is a quadratic programming problem because of the presence of terms such as X_i^2 and X_iX_j (squared and cross-product terms). However, there are standard packages available that solve this problem.

THE INCORPORATION OF ADDITIONAL CONSTRAINTS

The imposition of short sale constraints has complicated the solution technique, forcing us to use quadratic programming. Once we resort to this technique, however, it is a simple matter to impose other requirements on the solution. Literally any set of requirements that can be formulated as linear functions of the investment weights can be imposed on the solution. For example, some managers wish to select optimum portfolios given that the dividend yield on the optimum portfolios is at least some number (e.g., 2%). If we let D stand for the target dividend yield and d_i stand for the dividend yield on stock i, then we can impose this requirement by adding a fourth constraint to the problem described in the previous section:

$$(4)\quad \sum_{i=1}^{N}\left(X_id_i\right) \geq D$$

If we desire the dividend constraint but want to allow short sales, we simply eliminate the third constraint,

$$(3)\quad X_i \geq 0, \qquad i = 1,\ldots,N$$

from the problem.

Note that once we impose inequality constraints such as the one described for dividends, we must solve a quadratic programming problem instead of a system of simultaneous equations, even if short sales are allowed.

Other types of constraints are frequently employed in solving portfolio problems. Perhaps the most frequent constraints are those that place an upper limit on the fraction of the portfolio that can be invested in any stock. Upper limits on the amount that can be

Table 6.1 Input Data for Asset Allocation

	S&P	Bonds	Canadian	Japan	Emerging Market	Pacific	Europe	Small Stock
Expected return	14.00	6.50	11.00	14.00	16.00	18.00	12.00	17.00
Standard deviation	18.50	5.00	16.00	23.00	30.00	26.00	20.00	24.00
				Correlation Coefficients				
S&P	1.00	0.45	0.70	0.20	0.64	0.30	0.61	0.79
Bonds		1.00	0.27	−0.01	0.41	0.01	0.13	0.28
Canadian			1.00	0.14	0.51	0.29	0.48	0.59
Japan				1.00	0.25	0.73	0.56	0.13
Emerging market					1.00	0.28	0.61	0.75
Pacific						1.00	0.54	0.16
Europe							1.00	0.44
Small stock								1.00

invested in any one stock are often part of the charter of mutual funds. Also, upper limits (and occasionally lower limits) are often placed on the fraction of a portfolio that can be invested in any industry. Finally, it is possible to build in constraints on the amount of turnover in a portfolio and to allow the consideration of transaction costs in computing returns.

AN EXAMPLE

This chapter has presented techniques for obtaining the efficient frontier when there are a large number of assets to choose from. Table 6.1 shows the data for the asset allocation problem we will examine. The manager is considering an allocation across three U.S. categories and international stocks. The three U.S. categories are large stocks, small stocks, and bonds. Large stocks are represented by the Standard and Poor's index including dividends, bonds by Barclays Government Credit index, and small stocks by the Center for Research in Security Prices (CRSP) small stock index.[8] The international data were obtained by using returns on international stock mutual funds. The international portfolios are selected to divide the world into as many nonoverlapping segments as possible. Thus there is a Canadian fund, a European fund, a Japanese fund, a Pacific funds, and an emerging market fund. There is some overlap. The Pacific fund and the Japanese fund have stocks in Japan in common. Similarly, the emerging market and Pacific funds have some countries in common. The effect of overlap can be seen by examining the correlation coefficients. The correlation between the Japan fund and the Pacific fund is 0.73, which is the highest correlation between Japan and any other fund. The emerging market is interesting. Before examining the data, one would expect that the correlations would be very low with the major countries. However, the correlations are high with major markets, implying that the performance of emerging markets is very much affected by what happens in major markets.

The correlation matrix initially was calculated by using return data over the prior five years and was calculated for returns expressed in U.S. dollars. Then, security analysts at a

[8]The CRSP small stock index is roughly the smallest quintile of stocks on the New York Stock Exchange plus American Stock Exchange and NASDAQ stocks of similar size. See the footnote to Table 17.1 for a detailed description of the construction of the CRSP small stock index.

major investment banking firm compared the correlations calculated using returns from the most recent five-year period with prior five-year periods. Using these data and their judgment, analysts modified some historic numbers to obtain their best estimate of what the future correlations would be.

The standard deviations are expressed in annual returns. They were also calculated over the prior five years. Once again, however, analysts modified them slightly utilizing both data from earlier periods and their experience to obtain their best subjective estimates for the future. The mean returns are the estimates of a major financial intermediary concerned with the allocation decisions analyzed here. At this time they were fairly pessimistic about U.S. bond markets, Canadian stocks, and European stocks, and this is reflected in their estimates. The final input needed is a riskless rate of interest, which was estimated at 5% for U.S. investors over subsequent years.

The efficient frontier without riskless lending and borrowing but with short sales is the curved figure shown in Figure 6.4. Each asset class as a separate investment is represented by a dot in Figure 6.4. The global minimum variance portfolio has a mean return of 6.41% and a standard deviation of 3.91%. Note that bonds are by far the least risky asset. However, a portfolio of assets is less risky than bonds, even though the next least risky asset has a standard deviation more than 3 times larger than bonds. Alternatively, the optimum portfolio with the same risk as bonds has a mean return of 8.42%, or 1.92% more than bonds. This is an illustration of the power of diversification. Note that all assets are held either long or short. Furthermore, note that for the higher returns (above portfolio 2), the short sales involved are substantial and would involve short selling more than margin requirements would allow. Thus the efficient frontier would terminate after portfolio 2. At low risks, the major long purchase is bonds. As expected return is increased, the S&P, small stocks, and the Pacific fund all are held long in substantial amounts, with Japan held long in a somewhat smaller proportion. These are all relatively

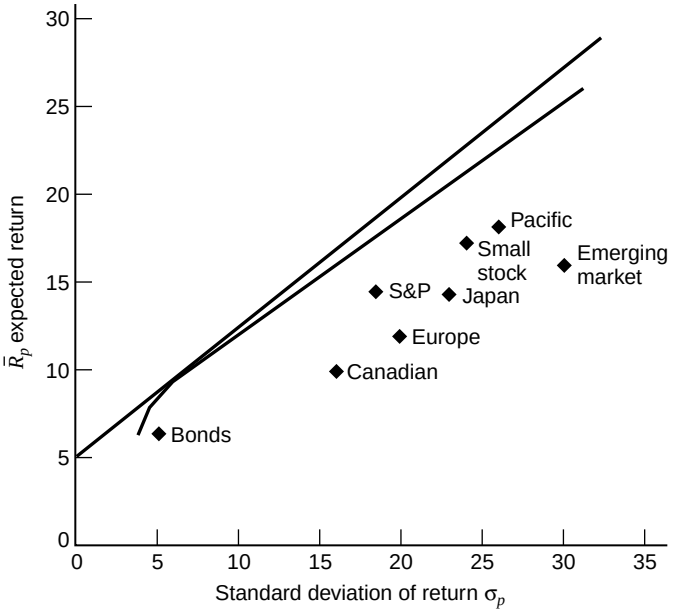


Figure 6.4 The efficient frontier with riskless lending and borrowing and short sales allowed.

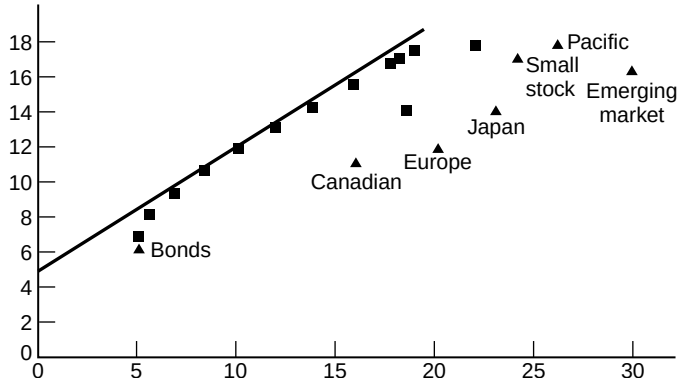


Figure 6.5 The efficient frontier with no riskless lending and borrowing and no short sales.

high-expected-return portfolios. Notice, however, that emerging markets, the other high mean return portfolio, does not enter into the optimum. This is because it has a very high correlation with the other countries and thus does not contribute much to the diversification. Europe and bonds are sold short for portfolios with higher mean returns. These are both low-expected-return assets. In addition, Europe has the advantage of being relatively highly correlated with the assets held long. When an asset is sold short, the covariance term with a long asset is negative, thus reducing risk. It is therefore desirable for a short-sold asset to be highly correlated with an asset held long.

Now consider the solution when short sales are not allowed and there is no riskless lending and borrowing. The efficient frontier is the curved region in Figure 6.5. The composition for a number of portfolios is shown in Table 6.2. The case where short sales are not allowed is probably the realistic case to consider for the pension fund manager whose problem we are analyzing. As shown in Table 6.2, the global minimum variance portfolio has an expected return of 6.89% and a standard deviation of 4.87%. This is of course a higher standard deviation than if short sales were allowed. A comparison of the numbers in Figures 6.1 and 6.2 shows that the efficient frontier with short sales allowed offers a higher mean return for a given risk (either with or without riskless lending and borrowing). This is because short sales offer additional investment opportunities that are used.

As shown in Table 6.2, the minimum-risk portfolio is primarily investment in bonds. Without short sales, the minimum risk is only slightly less than the risk of bonds alone—4.87% compared to 5%—and the expected return is only 0.39% higher. As we increase the risk on the portfolio, the percentage invested in bonds goes down, and we start to invest primarily in small stocks and Pacific. A minor amount is invested in Japan. The highest mean return portfolio is of course 100% in the highest-return asset, Pacific bond.

When riskless lending and borrowing is allowed, the efficient frontier is the straight line shown in Figures 6.4 and 6.5. The equations of the straight lines are

Short sales allowed

$$\overline{R}_P = 5 + 0.714\sigma_P$$

Short sales not allowed

$$\overline{R}_P = 5 + 0.685\sigma_P$$

Table 6.2 Proportions Invested When Short Sales Are Not Allowed

	Global Minimum	1	2	3	4	5
Mean return	6.89	9.36	11.83	14.30	16.77	18.00
Standard deviation	4.88	6.66	10.03	13.86	17.87	26.00
			Proportions			
S&P	0.00	0.00	0.00	0.00	0.63	0.00
Bond	95.16	72.91	50.51	28.12	5.51	0.00
Canadian	0.06	0.00	0.00	0.00	0.00	0.00
Japan	3.96	3.57	3.17	2.77	2.41	0.00
Emerging market	0.00	0.00	0.00	0.00	0.00	0.00
Pacific	0.81	12.42	22.86	33.29	43.62	100.00
Europe	0.00	0.00	0.00	0.00	0.00	0.00
Small stock	0.00	11.10	23.46	35.82	47.82	0.00

Obviously, the efficient frontier with short sales allowed is steeper. The tangency portfolio for short sales not allowed has a mean return of 11.51%. Higher returns involve borrowing at the riskless rate. For the pension manager whose problem is being analyzed, this is likely infeasible. For this manager, the efficient frontier is likely to be the straight line segment from R_F to the tangency point and the curved shape from there to the right. Given the low return of the tangency portfolio, the choice would likely lie on the curve to the right of the tangency portfolio. This would involve bonds, small stocks, Pacific, and a little invested in Japan. It would be important to vary the inputs in a reasonable range to see how this composition would change given reasonable changes in the inputs.

CONCLUSION

In this chapter we discussed and illustrated the use of techniques that can be employed to solve for the set of all possible portfolios that are efficient. All of the solution techniques discussed are feasible and have been used to solve problems. However, the techniques require gigantic amounts of input data and large amounts of computation time. Furthermore, the input data are in a form to which the security analyst and portfolio manager cannot easily relate. For this reason, it is difficult to get estimates of the input data or to get practitioners to relate to the final output.

The next logical step is to simplify the number and type of input requirements for portfolio selection and, in turn, to see if this reduction in data complexity can also be used to simplify the computational procedure. This is the subject of the next three chapters.

APPENDIX A

AN ALTERNATIVE DEFINITION OF SHORT SALES

Modeling short sales from the viewpoint of the broker-dealer, we first note that the broker-dealer has a fixed sum of money to invest. A short sale involves putting up an amount of money equal to the short sale. Thus the short sale is a use rather than a source of funds to the short seller. The total funds the broker-dealer invests short, plus the funds invested long, must add to the original investment. Because for short sales, $X_i < 0$, the proportion of the funds invested in the short sale is $|X|_i$. In addition, the short seller (if a broker-dealer)

receives interest on both the money put up against short sales and the money received from the short sale. Thus the expected return from short selling 0.10 of stock i is $-0.1\bar{R}_i + 0.2R_F$. Because X_i is negative for short sales, this can be written as $X_i(\bar{R}_i - 2R_F)$. Assume stocks 1 to k are held long and stocks $k + 1$ to N are sold short. Then

$$\bar{R}_P = \sum_{i=1}^{k} X_i\left(\bar{R}_i\right) + \sum_{i=k+1}^{N} X_i\left(\bar{R}_i - 2R_F\right)$$

$$\bar{R}_P = \sum_{i=1}^{N} X_i\bar{R}_i - 2\sum_{i=k+1}^{N} X_i\left(R_F\right)$$

The constraint with the Lintner definition of short sales is

$$\sum_{i=1}^{N} |X| = 1$$

Substituting this for 1 times R_F yields

$$R_F = \sum_{i=1}^{N} |X_i| R_F = \sum_{i=1}^{k} X_i R_F - \sum_{i=k+1}^{N} X_i R_F \tag{A.1}$$

This is the expression used for R_F. Subtracting R_F from both sides of the equation for $\bar{R}_P$ and using (A.1) for R_F on the right-hand side of the equation yields

$$\bar{R}_P - R_F = \left[\sum_{i=1}^{N} X_i\bar{R}_i - 2\sum_{i=k+1}^{N} X_i R_F\right] - \left[\sum_{i=1}^{k} X_i R_F - \sum_{i=k+1}^{N} X_i R_F\right]$$

$$\bar{R}_P - R_F = \sum_{i=1}^{N} X_i\bar{R}_i - \sum_{i=1}^{N} X_i R_F = \sum_{i=1}^{N} X_i\left(\bar{R}_i - R_F\right)$$

This is identical to the equation given in the text. The reader should note that in the Lintnerian definition of short sales, the final portfolio weights must be scaled so that the sum of the absolute value of the weights, rather than their sum, is 1.

APPENDIX B

DETERMINING THE DERIVATIVE

In the text we discussed that to solve the portfolio problem when short sales are allowed, the derivative of θ with respect to X_k was needed.[9] In the text we presented the value of the derivative. In this appendix we derive its value. To determine the derivative, rewrite the θ shown in the text as

$$\theta = \left[\sum_{i=1}^{N} X_i\left(\bar{R}_i - R_F\right)\right]\left[\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1 \\ j\neq i}}^{N} X_i X_j\sigma_{ij}\right]^{-1/2}$$

[9]To ensure a maximum, the second derivative should be negative. The structure of this problem guarantees this.

Two rules from calculus are needed:

1. **The product rule:** θ is the product of two functions. The product rule states that the derivative of the product of two functions is the first function times the derivative of the second function plus the second times the derivative of the first. In symbols,

$$\frac{d}{dX}\left[\left[F_1(X)\right]\left[F_2(X)\right]\right] = \left[F_1(X)\right]\frac{dF_2(X)}{dX} + \left[F_2(X)\right]\frac{dF_1(X)}{dX} \tag{B.1}$$

Let

$$F_1(X) = \sum_{i=1}^{N} X_i\left(\overline{R}_i - R_F\right) \tag{B.2}$$

$$F_2(X) = \left(\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}\right)^{-1/2} \tag{B.3}$$

Consider first the derivative of $F_1(X)$. At first glance, the reader may believe it is difficult. However, it turns out to be trivial. An expression like

$$\sum_{i=1}^{N} X_i\left(\overline{R}_i - R_F\right)$$

involves a lot of terms that do not contain an X_k and one term involving X_k. The derivatives of the terms not involving X_k are zero (they are constants as far as X_k is concerned). The derivative of the term involving X_k is $\overline{R}_k - R_F$. Thus

$$\frac{dF_1(X)}{dX_k} = \overline{R}_k - R_F \tag{B.4}$$

Now consider the derivative of $F_2(X)$. To determine this, a second rule from calculus is needed.

2. **The chain rule:** $F_2(X)$ involves a term in brackets to a power (the power $-\frac{1}{2}$). The chain rule states that its derivative is the power, times the expression in parentheses to the power minus one, times the derivative of what is inside the brackets. Thus

$$\frac{dF_2(X)}{dX_k} = \left(-\frac{1}{2}\right)\left(\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}\right)^{-3/2} \times \left(2X_k\sigma_k^2 + 2\sum_{\substack{j=1\\ j\neq k}}^{N} X_j\sigma_{jk}\right) \tag{B.5}$$

The only term that requires comment is the last one. The derivative of

$$\sum_{i=1}^{N} X_i^2\sigma_i^2$$

follows the same principles discussed earlier. All terms not involving k are constant as far as k is concerned, and thus their derivative is zero. The term involving k is $X_k^2\sigma_k^2$ and has a derivative of $2X_k\sigma_k^2$. The derivation of the double summation is more complex. Consider the double summation term

$$\left(\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \sigma_{ij} \right)$$

We get X_k twice, once when $i = k$ and once when $j = k$. When $i = k$, we have

$$\sum_{\substack{j=1 \\ j \neq k}}^{N} X_k X_j \sigma_{kj} = X_k \left[\sum_{\substack{j=1 \\ j \neq k}}^{N} X_j \sigma_{kj} \right]$$

The derivative of this is, of course,

$$\sum_{\substack{j=1 \\ j \neq k}}^{N} X_j \sigma_{kj}$$

Similarly, when $j = k$, we have

$$\sum_{\substack{i=1 \\ i \neq k}}^{N} X_i X_k \sigma_{ik} = X_k \left(\sum_{\substack{i=1 \\ i \neq k}}^{N} X_i \sigma_{ik} \right)$$

The derivative of this is also

$$\sum_{\substack{i=1 \\ i \neq k}}^{N} X_i \sigma_{ik}$$

where i and j are simply summation indicators. It does not matter which we use. Furthermore, $\sigma_{ik} = \sigma_{ki}$. Thus

$$\sum_{\substack{j=1 \\ j \neq k}}^{N} X_j \sigma_{kj} = \sum_{\substack{i=1 \\ i \neq k}}^{N} X_i \sigma_{ik}$$

and we have the expression shown in the derivative, namely,

$$2 \sum_{\substack{j=1 \\ j \neq k}}^{N} X_j \sigma_{kj}$$

Substituting (B.2), (B.3), (B.4), and (B.5) into the product rule, expression (B.1) yields

$$\frac{d\theta}{dX_k} = \left[\sum_{i=1}^{N} X_i\left(\overline{R}_i - R_F\right)\right]\left[\left(-\frac{1}{2}\right)\left(\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}\right)^{-3/2}\right.$$
$$\left.\times\left(2X_k\sigma_k^2 + 2\sum_{\substack{j=1\\ j\neq k}}^{N} X_j\sigma_{kj}\right)\right] + \left[\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}\right]^{-1/2}$$
$$\times\left[\left(\overline{R}_k - R_F\right)\right] = 0$$

Multiplying the derivative by

$$\left(\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}\right)^{1/2}$$

and rearranging yields

$$-\left[\frac{\sum_{i=1}^{N} X_i\left(\overline{R}_i - R_F\right)}{\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}}\right]\left[X_k\sigma_k^2 + \sum_{\substack{j=1\\ j\neq k}}^{N} X_j\sigma_{kj}\right] + \left[\overline{R}_k - R_F\right] = 0$$

Defining λ as

$$\frac{\sum_{i=1}^{N} X_i\left(\overline{R}_i - R_F\right)}{\sum_{i=1}^{N} X_i^2\sigma_i^2 + \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} X_i X_j \sigma_{ij}}$$

yields

$$-\lambda\left[X_k\sigma_k^2 + \sum_{\substack{j=1\\ j\neq k}}^{N} X_j\sigma_{kj}\right] + \left(\overline{R}_k - R_F\right) = 0$$

Multiplying the terms in the brackets by λ yields

$$-\left[\lambda X_k \sigma_k^2 + \sum_{\substack{j=1 \\ j \neq k}}^{N} \lambda X_j \sigma_{kj}\right] + \left(\bar{R}_k - R_F\right) = 0$$

This is the expression shown in the text.

APPENDIX C

SOLVING SYSTEMS OF SIMULTANEOUS EQUATIONS

To solve large systems of simultaneous equations, one would use any of the large number of standard computer packages that exist for this purpose. However, small systems can be solved by hand. The simplest way is by repetitive substitution. Consider the following system of simultaneous equations:

$$4X_1 + X_2 = 7 \tag{C.1}$$

$$3X_1 + 2X_2 = 5 \tag{C.2}$$

Equation (C.1) can be rearranged so that X_2 is expressed as a function of X_1. This rearrangement yields

$$X_2 = 7 - 4X_1$$

Substituting this into Equation (C.2) yields

$$\begin{aligned} 3X_1 + 2(7 - 4X_1) &= 5 \\ 3X_1 + 14 - 8X_1 &= 5 \\ -5X_1 &= -9 \\ X_1 &= \tfrac{9}{5} \end{aligned}$$

Substituting the value for X_1 into rearranged Equation (C.1) yields

$$X_2 = 7 - 4\left(\tfrac{9}{5}\right) = 7 - \tfrac{36}{5} = -\tfrac{1}{5}$$

This technique is extremely easy and can be applied to solving any number of simultaneous equations, although with many equations, it becomes extremely time consuming. For a second example, consider the problem analyzed in the section "Short Sales Allowed":

$$1 = 4Z_1 + Z_2 + 2Z_3 \tag{C.3}$$

$$1 = 3Z_1 + 3Z_2 + 6Z_3 \tag{C.4}$$

$$5 = 6Z_1 + 6Z_2 + 75Z_3 \tag{C.5}$$

Solving Equation (C.3) for Z_2 and eliminating Z_2 from Equation (C.4) yields

$$Z_2 = 1 - 4Z_1 - 2Z_3 \tag{C.3′}$$

$$1 = 3Z_1 + 3(1 - 4Z_1 - 2Z_3) + 6Z_3 \tag{C.4′}$$

Simplifying (C.4′) yields

$$-2 = -9Z_1$$

Following the same procedure for Equation (C.5) yields

$$5 = 6Z_1 + 6(1 - 4Z_1 - 2Z_3) + 75Z_3 \qquad (C.5')$$

Simplifying (C.5′) yields

$$-1 = -18Z_1 + 63Z_3$$

Equation (C.4′) gives an immediate solution for Z_1; it is $Z_1 = \frac{2}{9}$. Substituting this into Equation (C.5′) allows us to solve for Z_3:

$$-1 = -18\left(\frac{2}{9}\right) + 63Z_3$$

$$-1 = -4 + 63Z_3$$

$$Z_3 = \frac{3}{63}$$

Substituting the values of Z_3 and Z_1 into (C.3′) yields for Z_2

$$Z_2 = 1 - \frac{8}{9} - \frac{6}{63} = \frac{1}{63}$$

This is the solution stated in the text. When the number of equations and variables becomes large, it is usually more convenient to solve the problem by working on a tableau. A tableau for the last problem is presented here.

Z_1	Z_2	Z_3	= Constant
4	1	2	= 1
3	3	6	= 1
6	6	75	= 5

Under each of the variables is the coefficient shown in the system of Equations (C.3), (C.4), and (C.5). If c_1, c_2, c_3 are arbitrary constants, the solution is reached when the tableau looks as follows:

Z_1	Z_2	Z_3	= Constant
1	0	0	c_1
0	1	0	c_2
0	0	1	c_3

To move from the first tableau to the second, three operations are allowed:

1. You can multiply or divide any row by a constant.

2. You can add or subtract a constant times one row from another row.

3. You can exchange any two rows.

Let us apply this to the problem discussed earlier. Subtracting twice row 2 from row 3 yields

Z_1	Z_2	Z_3	= Constant
4	1	2	1
3	3	6	1
0	0	63	3

Dividing row 3 by 63 yields

Z_1	Z_2	Z_3	= Constant
4	1	2	1
3	3	6	1
0	0	1	$\frac{3}{63}$

Subtracting 2 times row 3 from row 1, and 6 times row 3 from row 2 yields

Z_1	Z_2	Z_3	= Constant
4	1	0	$\frac{57}{63}$
3	3	0	$\frac{45}{63}$
0	0	1	$\frac{3}{63}$

Subtracting $\frac{1}{3}$ of row 2 from row 1 yields

Z_1	Z_2	Z_3	= Constant
3	0	0	$\frac{42}{63}$
3	3	0	$\frac{45}{63}$
0	0	1	$\frac{3}{63}$

Taking $\frac{1}{3}$ of row 1 and $\frac{1}{3}$ of row 2 yields

Z_1	Z_2	Z_3	= Constant
1	0	0	$\frac{14}{63}$
1	1	0	$\frac{15}{63}$
0	0	1	$\frac{3}{63}$

Subtracting row 1 from row 2 yields the final tableau:

Z_1	Z_2	Z_3	= Constant
1	0	0	$\frac{14}{63}$
0	1	0	$\frac{1}{63}$
0	0	1	$\frac{3}{63}$

The now familiar solution can be read directly from this tableau. It is

$$Z_1 = \frac{14}{63}, \qquad Z_2 = \frac{1}{63}, \qquad \text{and} \qquad Z_3 = \frac{3}{63}$$

Either of these methods can be used to solve a system of simultaneous equations.

APPENDIX D

A GENERAL SOLUTION

Although we have just outlined a feasible procedure for identifying the efficient frontier, there is a simpler one. When we assumed a particular riskless lending and borrowing rate, we determined that the optimum portfolio is the one that solves the following system of simultaneous equations:

$$\begin{aligned}
\bar{R}_1 - R_F &= Z_1\sigma_1^2 + Z_2\sigma_{12} + Z_3\sigma_{13} + \cdots + Z_N\sigma_{1N} \\
\bar{R}_2 - R_F &= Z_1\sigma_{12} + Z_2\sigma_2^2 + Z_3\sigma_{23} + \cdots + Z_N\sigma_{2N} \\
\bar{R}_3 - R_F &= Z_1\sigma_{13} + Z_2\sigma_{23} + Z_3\sigma_3^2 + \cdots + Z_N\sigma_{3N} \\
&\vdots \\
\bar{R}_N - R_F &= Z_1\sigma_{1N} + Z_2\sigma_{2N} + Z_3\sigma_{3N} + \cdots + Z_N\sigma_{N}
\end{aligned}$$

When we solved this system of simultaneous equations, we substituted, in particular, values of $\bar{R}_i$, R_F, σ_i^2, and σ_{ij}. However, we do not have to substitute in a particular value of R_F. We can simply leave R_F as a general parameter and solve for Z_k in terms of R_F. This results in a solution of the form

$$Z_k = C_{0k} + C_{1k}R_F$$

where C_{0k} and C_{1k} are constants. They have a different value for each security k, but that value does not change with changes in R_F. Once the Z_k are determined as functions of R_F, we could vary R_F to determine the amount to invest in each security at various points along the efficient frontier. Let us apply this to the example following Equation (6.1). The system of simultaneous equations for a general R_F is

$$14 - R_F = 36Z_1 + 9Z_2 + 18Z_3 \tag{D.1}$$

$$8 - R_F = 9Z_1 + 9Z_2 + 18Z_3 \tag{D.2}$$

$$20 - R_F = 18Z_1 + 18Z_2 + 225Z_3 \tag{D.3}$$

The solution to this system of simultaneous equations is

$$Z_1 = \frac{42}{189} \tag{D.4}$$

$$Z_2 = \frac{118}{189} - \frac{23}{189}R_F \tag{D.5}$$

$$Z_3 = \frac{4}{189} + \frac{1}{189}R_F \tag{D.6}$$

This solution can be confirmed by substituting these values into Equations (D.1), (D.2), and (D.3). Also, as a further check, note that the substitution of $R_F = 5$ (which was the value we assumed in the last section) into Equations (D.4), (D.5), and (D.6) yields

$$Z_1 = \frac{42}{189} = \frac{14}{63}$$
$$Z_2 = \frac{118}{189} - \frac{23}{189}(5) = \frac{118-115}{189} = \frac{3}{189} = \frac{1}{63}$$
$$Z_3 = \frac{4}{189} + \frac{1}{189}(5) = \frac{9}{189} = \frac{3}{63}$$

the same solution we obtained earlier. The values of Z_k just determined can be scaled to sum to 1 exactly as was done before so that the optimum proportions can be determined.

Determining the General Coefficient from Two Portfolios

In the last section we determined that

$$Z_2 = \frac{118}{189} - \frac{23}{189} R_F$$

Assume that we had not determined this general expression. Rather, we simply solved the system of simultaneous equations for two arbitrary values of R_F. The value of Z_2 corresponding to an R_F of 5 is $\frac{1}{63}$, and the Z_2 corresponding to an R_F of 2 is $\frac{72}{189}$. Can we determine the general expression? The answer is clearly yes. As an example, assume we had solved the equations for an R_F of 2 and 5. We know the general expression has the form

$$Z_2 = C_{02} + C_{12} R_F$$

Furthermore, we know that

$$Z_2 = \tfrac{1}{63} \qquad \text{if } R_F = 5$$
$$Z_2 = \tfrac{72}{189} \qquad \text{if } R_F = 2$$

Utilizing this in the previous equation, we have

$$\tfrac{1}{63} = C_{02} + C_{12}(5)$$
$$\tfrac{72}{189} = C_{02} + C_{12}(2)$$

This is a system of two equations and two unknowns. We can use it to solve for $C_{02} = \frac{118}{189}$ and $C_{12} = -\frac{23}{189}$. Thus, if we have the optimum portfolio for any two values of R_F, we can obtain the value for C_{0k} and C_{1k} and then, by varying R_F, obtain the full efficient frontier.

This is an extremely powerful result. It means that the solution of the system of simultaneous equations for any two values of R_F allows us to trace out the full efficient frontier.

The tracing out of the efficient frontier can be done in two ways. First, we could solve for the general expression for Z_k in terms of R_F by determining Z_k for any two arbitrary values of R_F. Then, by varying R_F over the relevant range, we could trace out the efficient frontier.

A second procedure is suggested by the previous discussion. We showed that solving the system of simultaneous equations for any two values of R_F allowed us to obtain a general

expression for Z_k in terms of R_F, thus enabling us to trace out the efficient frontier. This suggests that the efficient frontier can be determined directly simply by calculating any two optimum portfolios rather than indirectly by first determining Z_k as a function of R_F. It can be shown that this direct procedure is appropriate.[10] Thus the entire efficient frontier can be traced out by determining the composition of any two portfolios and then determining all combinations of these two portfolios. This is an extremely powerful result and is the preferred way to determine the efficient set.

In the previous chapter we showed how to trace out all combinations (portfolios) of two assets. Nothing prevents the two assets from being efficient portfolios. Thus, given that the efficient frontier can be traced out by combining two efficient portfolios, if we find two efficient portfolios, we can utilize the procedures discussed in the last chapter to trace out the full efficient frontier. Let us see how this is done.

Tracing Out the Efficient Frontier

The Z_k that correspond to an $R_F = 2$ are from Equations (D.4), (D.5), and (D.6):

$$Z_1 = \frac{42}{189}, \qquad Z_2 = \frac{72}{189}, \qquad Z_3 = \frac{6}{189}$$

The proportions to invest in each security are

$$X_1 = \frac{42}{120} = \frac{7}{20}$$
$$X_2 = \frac{72}{120} = \frac{12}{20}$$
$$X_3 = \frac{6}{120} = \frac{1}{20}$$

The expected return associated with this portfolio is

$$\bar{R}_P = \left(\frac{7}{20}\right)(14) + \left(\frac{12}{20}\right)(8) + \left(\frac{1}{20}\right)(20) = 10\tfrac{7}{10}$$

The variance of return on this portfolio is

$$\sigma_P^2 = \left(\frac{7}{20}\right)^2 (36) + \left(\frac{12}{20}\right)^2 9 + \left(\frac{1}{20}\right)^2 (225)$$
$$+2\left(\frac{7}{20}\right)\left(\frac{12}{20}\right)9 + 2\left(\frac{7}{20}\right)\left(\frac{1}{20}\right)(18) + 2\left(\frac{12}{20}\right)\left(\frac{1}{20}\right)18 = \frac{5{,}481}{400}$$

If we knew the covariance between the portfolios associated with an $R_F = 5$ and an $R_F = 2$, we could trace out the full efficient frontier by treating each portfolio as an asset and utilizing the method discussed in Chapter 5. The covariance is determined as follows.

[10] See Black (1972) for a rigorous proof that this holds.

Consider a portfolio consisting of $\frac{1}{2}$ of each of the two portfolios already determined. The investment proportions are

$$X_1'' = \frac{1}{2}\frac{7}{20} + \frac{1}{2}\frac{14}{18} = \frac{203}{360}$$
$$X_2'' = \frac{1}{2}\frac{12}{20} + \frac{1}{2}\frac{1}{18} = \frac{118}{360}$$
$$X_3'' = \frac{1}{2}\frac{1}{20} + \frac{1}{2}\frac{3}{18} = \frac{39}{360}$$

Its variance is

$$\sigma_P^2 = \left(\frac{203}{360}\right)^2 36 + \left(\frac{118}{360}\right)^2 9 + \left(\frac{39}{360}\right)^2 225$$
$$+ 2\left(\frac{203}{360}\right)\left(\frac{118}{360}\right)9 + 2\left(\frac{203}{360}\right)\left(\frac{118}{360}\right)18$$
$$+ 2\left(\frac{118}{360}\right)\left(\frac{39}{360}\right)18 = 21.859$$

But we know that this portfolio is a weighted average of the other two portfolios. In Chapter 5 we showed that the variance of a portfolio composed of two assets or portfolios was

$$\sigma_P^2 + X_1^2\sigma_1^2 + X_2^2\sigma_2^2 + 2X_1X_2\sigma_{12}$$

Thus the variance of a portfolio consisting of $\frac{1}{2}$ of portfolio 1 and $\frac{1}{2}$ of portfolio 2 is

$$\sigma^2 = \left(\frac{1}{2}\right)^2\left(\frac{203}{6}\right) + \left(\frac{1}{2}\right)^2\left(\frac{5481}{400}\right) + 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\sigma_{12}$$

We know the variance of this portfolio is 21.859. Thus σ_{12} can be determined from

$$21.859 = \left(\frac{1}{2}\right)^2\left(\frac{203}{6}\right) + \left(\frac{1}{2}\right)^2\left(\frac{5481}{400}\right) + 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\sigma_{12}$$

and

$$\sigma_{12} = 19.95$$

Knowing the expected return variance and covariance, we can trace out the efficient frontier exactly as we did for combinations of two assets in Chapter 5. We have done so in Figure 6.6.

The Number of Securities Included

Before leaving this section, some observations are in order. First, when short sales are allowed, the investor takes a position in almost all securities. Each security will have, in general, one value of R_F for which it is not held, namely, when $C_{0k} + C_{1k}R_F = 0$. But for all other values of R_F, it will be held either long or short. In fact, for all values of R_F above this value, the security will be held only long or short, and vice versa for values of R_F

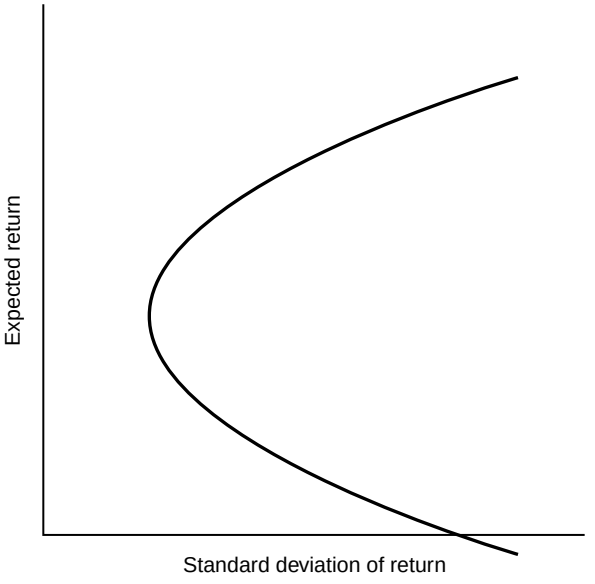


Figure 6.6 The minimum variance frontier.

below the value. Let us examine the expressions for Z_k as a function of R_F from our previous example:

$$Z_1 = \frac{42}{189}$$
$$Z_2 = \frac{118}{189} - \frac{23}{189} R_F$$
$$Z_3 = \frac{4}{189} + \frac{1}{189} R_F$$

Security 1 is always held long. Security 2 is held long if R_F is less than $\frac{118}{23}$ and short for all values of R_F greater than $\frac{118}{23}$. Finally, security 3 is held long if R_F is greater than -4 and short for values of R_F below -4. The various values of Z as a function of R_F are shown in Figure 6.7.

The inclusion of almost all or all securities in the optimum portfolio makes intuitive sense. If a security's characteristics make it undesirable to hold, then the investor should issue it by selling it short. Thus "good" securities are held and "bad" securities are issued to someone else. Of course, for someone else to be willing to take "bad" securities, there has to be a difference of opinion regarding what is good and what is bad.

APPENDIX E

QUADRATIC PROGRAMMING AND KUHN–TUCKER CONDITIONS

These quadratic programming algorithms are based on a technique from advanced calculus called Kuhn–Tucker conditions. For small-scale problems, these conditions may be able to be used directly. Furthermore, an understanding of the nature of the solution to this type of portfolio problem can be gained by understanding the Kuhn–Tucker conditions.

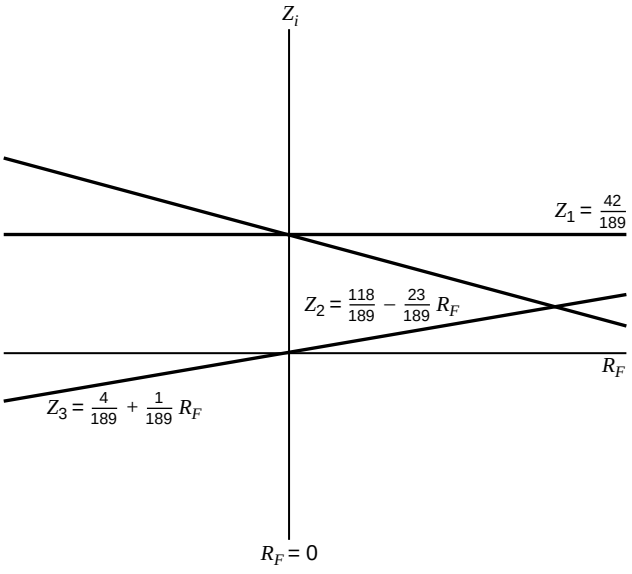


Figure 6.7 Portfolio proportion as a function of the riskless rate.

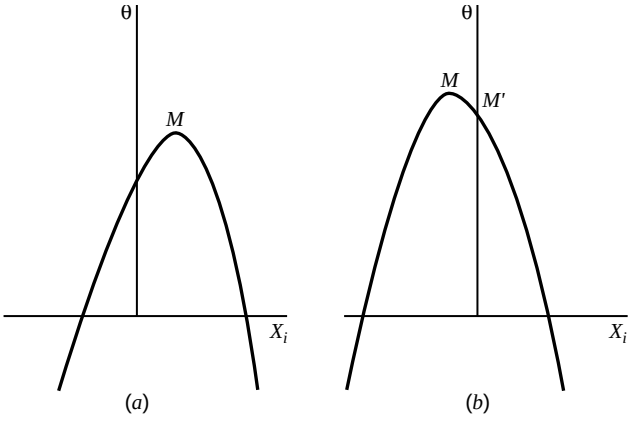


Figure 6.8 Value of the function as X changes.

Earlier we simply took the derivative of θ with respect to each X_i and set it equal to zero to find a maximum value of θ. This maximum is indicated by point M in Figure 6.8*a* or 6.8*b*. When X_i must be nonnegative, a problem can occur because the unconstrained maximum may be at a value of X_i, which is infeasible. Variable θ as a function of X_i might look like Figure 6.8*b* rather than Figure 6.8*a*. In this case (Figure 6.8*b*), the maximum feasible value of θ occurs at point M' rather than M. Notice that if the maximum value for X_i occurs at M', then $d\theta/dX_i < 0$ at the maximum feasible value ($X_i = 0$), whereas if it occurs when X_i is positive, then $d\theta/dX_i = 0$. Thus, in general, with X_i constrained to be larger than or equal to zero, we can write

$$\frac{d\theta}{dX_i} \leq 0$$

We could make this an equality by writing

$$\frac{d\theta}{dX_i} + U_i = 0$$

This is the first Kuhn–Tucker condition for a maximum.

Note two things about U_i. If the optimum occurs when X_i is positive, then the $d\theta/dX_i = 0$ and U_i is zero. Furthermore, if the optimum occurs when the maximum occurs at $X_i = 0$, then $d\theta/dX_i < 0$ and U_i is positive. To summarize, at the optimum we have

$$\begin{aligned} X_i > 0, &\qquad U_i = 0 \\ X_i = 0, &\qquad U_i > 0 \end{aligned}$$

This is the second Kuhn–Tucker condition. It can be written compactly as

$$\begin{aligned} X_i U_i &= 0 \\ X_i &\geq 0 \\ U_i &\geq 0 \end{aligned}$$

The four Kuhn–Tucker conditions are

$$\begin{aligned} &(1) \quad \frac{d\theta}{dX_i} + U_i = 0 \\ &(2) \quad X_i U_i = 0 \\ &(3) \quad X_i \geq 0 \\ &(4) \quad U_i \geq 0 \end{aligned}$$

If someone suggested a solution to us and it satisfied the Kuhn–Tucker conditions, then we could be sure that he had indeed given us the optimum portfolio.[11] For example, assume the lending and borrowing rate was 6% and the securities being considered are the three securities considered throughout this chapter. Furthermore, assume the solution was

$$\begin{aligned} X_1 = \frac{43}{53}, &\qquad U_1 = 0 \\ X_2 = 0, &\qquad U_2 = \frac{5}{8} \\ X_3 = \frac{10}{53}, &\qquad U_3 = 0 \end{aligned}$$

Because this solution meets all the Kuhn–Tucker conditions, it is optimal.

[11]There are conditions on the shape of θ for this to be optimum, but they are always met for the portfolio problem and so can be safely ignored here.

To see that this solution meets the Kuhn–Tucker conditions, consider the following. First, all Xs and Us are positive; thus conditions 3 and 4 are met. U_1, X_2, and $U_3 = 0$; thus either X or U is zero for any pair of securities, and condition 2 is met. Finally, recall that

$$\frac{d\theta}{dX_i} = \bar{R}_i - R_F - \lambda\left[X_i\sigma_i^2 + \sum_{\substack{j=1 \\ j\neq i}}^{N} X_j\sigma_{ij}\right]$$

Adding U_i to this equation and substituting in the returns, variances, and covariances for the various securities, we have

$$\begin{aligned} &8 - \lambda\left[36X_1 + 9X_2 + 18X_3\right] + U_1 \\ &2 - \lambda\left[9X_1 + 9X_2 + 18X_3\right] + U_2 \\ &14 - \lambda\left[18X_1 + 18X_2 + 225X_3\right] + U_3 \end{aligned}$$

where $\lambda = (\bar{R}_P - R_F)/\sigma_P^2$. A little calculation shows that $\lambda = \frac{53}{216}$.

Substituting for X_1, X_2, and X_3 yields

$$\begin{aligned} &8 - \frac{53}{216}\left[36\left(\frac{43}{53}\right) + 9(0) + 18\left(\frac{10}{53}\right)\right] + 0 \\ &2 - \frac{53}{216}\left[9\left(\frac{43}{53}\right) + 9(0) + 18\left(\frac{10}{53}\right)\right] + \frac{5}{8} \\ &14 - \frac{53}{216}\left[18\left(\frac{43}{53}\right) + 18(0) + 225\left(\frac{10}{53}\right)\right] + 0 \end{aligned}$$

Because all three equal zero, the Kuhn–Tucker conditions are met.

QUESTIONS AND PROBLEMS

1. Assume analysts provide the following types of information. Assume (standard definition) short sales are allowed. What is the optimum portfolio if the lending and borrowing rate is 5%?

Security	Mean Return	Standard Deviation	Covariance with A	Covariance with B	Covariance with C
A	10	4		20	40
B	12	10			70
C	18	14			

2. Given the following information, what is the optimum portfolio if the lending and borrowing rate is 6%, 8%, or 10%? Assume the Lintner definition of short sales.

Security	Mean Return	Standard Deviation	Covariance with A	Covariance with B	Covariance with C
A	11	2		10	4
B	14	6			30
C	17	9			

3. Assume the information given in Problem 1 but that short sales are not allowed. Set up the formulation necessary to solve the portfolio problem.

4. Consider the following data. What is the optimum portfolio, assuming short sales are allowed (standard definition)? Trace out the efficient frontier.

Number	$\bar{R}_i$	σ_i
1	10	5
2	8	6
3	12	4
4	14	7
5	6	2
6	9	3
7	5	1
8	8	4
9	10	4
10	12	2
	$\rho_{ij} = 0.5$ for all ij	
	$R_F = 4$	

5. Assume that the data below apply to two efficient portfolios. What is the efficient frontier? Assume the standard definition of short sales.

Portfolio	$\bar{R}_i$	σ_i	
A	10	6	
B	8	4	$\sigma_{ij} = 20$

BIBLIOGRAPHY

1. Alexander, Gordon. "The Derivation of Efficient Sets," *Journal of Financial and Quantitative Analysis*, **XI,** No. 5 (Dec. 1976), pp. 817–830.
2. ——. "Mixed Security Testing of Alternative Portfolio Selection Modes," *Journal of Financial and Quantitative Analysis*, **XII,** No. 4 (Dec. 1977), pp. 817–832.
3. ——. "A Reevaluation of Alternative Portfolio Selection Models Applied to Common Stocks," *Journal of Financial and Quantitative Analysis*, **XIII,** No. 1 (March 1978), pp. 71–78.
4. Bawa, Vijay. "Mathematical Programming of Admissible Portfolios," *Management Science*, **23,** No. 7 (March 1977), pp. 779–785.
5. Bawa, Vijay S., Brown, Stephen J., and Klein, Roger W. *Estimation Risk and Optimal Portfolio Choice.* (Amsterdam: North Holland, 1979).
6. Bertsekas, Dimitris. "Necessary and Sufficient Conditions for Existence of an Optimal Portfolio," *Journal of Economic Theory*, **8,** No. 2 (June 1974), pp. 235–247.
7. Black, Fisher. "Capital Market Equilibrium with Restricted Borrowing," *Journal of Business*, **45,** No. 3 (July 1972), pp. 444–445.

8. Bowden, Roger. "A Dual Concept and Associated Algorithm in Mean–Variance Portfolio Analysis," *Management Science*, **23,** No. 4 (Dec. 1976), pp. 423–432.
9. Breen, William, and Jackson, Richard. "An Efficient Algorithm for Solving Large-Scale Portfolio Problems," *Journal of Financial and Quantitative Analysis*, **VI,** No. 1 (Jan. 1971), pp. 627–637.
10. Buser, Stephen. "Mean–Variance Portfolio Selection with Either a Singular or Non-Singular Variance–Covariance Matrix," *Journal of Financial and Quantitative Analysis*, **XII,** No. 3 (Sept. 1977), pp. 436–461.
11. Chen, Andrew. "Portfolio Selection with Stochastic Cash Demand," *Journal of Financial and Quantitative Analysis*, **XII,** No. 2 (June 1977), pp. 197–213.
12. Chen, Andrew, Jen, Frank, and Zionts, Stanley. "The Optimal Portfolio Revision Policy," *Journal of Business*, **44,** No. 1 (Jan. 1971), pp. 51–61.
13. ——. "Portfolio Models with Stochastic Cash Demands," *Management Science*, **19,** No. 3 (Nov. 1972), pp. 319–332.
14. Chen, Andrew, Kim, Han, and Kon, Stanley. "Cash Demands, Liquidation Costs and Capital Market Equilibrium under Uncertainty," *Journal of Financial Economics*, **2,** No. 3 (Sept. 1975), pp. 293–308.
15. ——. "Cash Demand...Reply," *Journal of Financial Economics*, **3,** No. 3 (June 1976), pp. 297–298.
16. Constantinides, George. "Comment on Chen, Kim and Kon," *Journal of Financial Economics*, **3,** No. 3 (June 1976), pp. 295–296.
17. Dybvig, Philip H. "Short Sales Restrictions and Kinks on the Mean Variance Frontier," *Journal of Finance*, **39,** No. 1 (March 1984), pp. 239–244.
18. Faaland, Bruce. "An Integer Programming Algorithm for Portfolio Selection," *Management Science*, **20,** No. 10 (June 1974), pp. 1376–1384.
19. Fishburn, Peter, and Porter, Burr. "Optimal Portfolios with One Safe and One Risky Asset: Effects of Change in Rate of Return and Risk," *Management Science*, **22,** No. 10 (June 1976), pp. 1064–1073.
20. Hill, Rowland. "An Algorithm for Counting the Number of Possible Portfolios Given Linear Restrictions on the Weights," *Journal of Financial Economics*, **XI,** No. 3 (Sept. 1976), pp. 479–487.
21. Jacob, Nancy. "A Limited-Diversification Portfolio Selection Model for the Small Investor," *Journal of Finance*, **XXIX,** No. 3 (June 1974), pp. 847–856.
22. Jones-Lee, M. W. "Some Portfolio Adjustment Theorems for the Case of Non-Negativity Conditions on Security Holdings," *Journal of Finance*, **XXVI,** No. 3 (June 1971), pp. 763–775.
23. Jorion, Philippe. "Bayes-Stein Estimation for Portfolio Analysis," *Journal of Financial and Quantitative Analysis*, **21,** No. 3 (Sept. 1986), pp. 279–292.
24. Lewis, Alan L. "A Simple Algorithm for the Portfolio Selection Problem," *Journal of Finance*, **43,** No. 1 (March 1988), pp. 71–82.
25. Lintner, John. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, **XLVII** (Feb. 1965), pp. 13–37.
26. Shanken, Jay. "Testing Portfolio Efficiency When the Zero-Beta Rate Is Unknown: A Note," *Journal of Finance*, **41,** No. 1 (March 1986), pp. 269–276.
27. Tucker, James, and Defaro, Clovis. "A Simple Algorithm for Stone's Version of the Portfolio Selection Problem," *Journal of Financial and Quantitative Analysis*, **X,** No. 5 (Dec. 1975), pp. 859–870.
28. Ziemba, William. "Solving Nonlinear Programming Problems with Stochastic Objective Functions," *Journal of Financial and Quantitative Analysis*, **VII,** No. 3 (June 1972), pp. 1809–1827.

Section 2

Simplifying the Portfolio Selection Process

7

The Correlation Structure of Security Returns—the Single-Index Model

In the previous three chapters of this book we outlined the basics of modern portfolio theory. The core of the theory, as described in these chapters, is not new; in fact, it was presented as early as 1956 in Markowitz's pioneering article and subsequent book. The reader, noting that the theory is over 50 years old, might well ask what has happened since the theory was developed. Furthermore, if you had knowledge about the actual practices of financial institutions, you might well ask why the theory took so long to be used by financial institutions. The answers to both these questions are closely related. Most of the research on portfolio management in the last 50 years has concentrated on methods for *implementing* the basic theory. Many of the breakthroughs in implementation have been quite recent, and it is only with these new contributions that portfolio theory becomes readily applicable to the management of actual portfolios.

In the next three chapters we are concerned with the implementation of portfolio theory. Breakthroughs in implementation fall into two categories: the first concerns a simplification of the amount and type of input data needed to perform portfolio analysis. The second involves a simplification of the computational procedure needed to calculate optimal portfolios. As will soon become clear, these issues are interdependent. Furthermore, their resolution vastly simplifies portfolio analysis. This results in the ability to describe the problem and its solution in relatively simple terms—terms that have intuitive as well as analytical meaning, and terms to which practicing security analysts and portfolio managers can relate.

In this chapter we begin the problem of simplifying the inputs to the portfolio problem. We start with a discussion of the amount and type of information needed to solve a portfolio problem. We then discuss the oldest and most widely used simplification of the portfolio structure: the single-index model. The nature of the model as well as some estimating techniques are examined.

In Chapter 8 we discuss alternative simplified representations of the portfolio problem. In particular, we are concerned with other ways to represent and predict the correlation structure between returns. Finally, in the last chapter, dealing with implementation, we show how each of the techniques that have been developed to simplify the input to portfolio analysis can be used to reduce and simplify the calculations needed to find optimal portfolios.

Most of Chapters 7 and 8 will be concerned with simplifying and predicting the correlation structure of returns. Many of the single- and multi-index models discussed in these chapters were developed to aid in portfolio management. Lately, however, these models have been used for other purposes that are often viewed as being as important as their use in portfolio analysis. Although many of these other uses will be detailed later in the book, we briefly describe some of them at the end of this chapter and in Chapter 8.

THE INPUTS TO PORTFOLIO ANALYSIS

Let us return to a consideration of the portfolio problem. From earlier chapters we know that to define the efficient frontier, we must be able to determine the expected return and standard deviation of return on a portfolio. We can write the expected return on any portfolio as

$$\bar{R}_P = \sum_{i=1}^{N} X_i \bar{R}_i \tag{7.1}$$

while the standard deviation of return on any portfolio can be written as

$$\sigma_P = \left[\sum_{i=1}^{N} X_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \sigma_i \sigma_j \rho_{ij} \right]^{1/2} \tag{7.2}$$

These equations define the input data necessary to perform portfolio analysis. From Equation (7.1) we see that we need estimates of the expected return on each security that is a candidate for inclusion in our portfolio. From Equation (7.2) we see that we need estimates of the variance of each security, plus estimates of the correlation between each possible pair of securities for the stocks under consideration. The need for estimates of correlation coefficients differs both in magnitude and substance from the two previous requirements. Let's see why.

The principal job of the security analyst traditionally has been to estimate the future performance of stocks he follows. At a minimum, this means producing estimates of expected returns on each stock he follows.[1]

With the increased attention that "risk" has received in recent years, more and more analysts are providing estimates of risk as well as return. The analyst who estimates the expected return of a stock should also be in a position to estimate the uncertainty of that return.

Correlations are an entirely different matter. Portfolio analysis calls for estimates of the pairwise correlation between all stocks that are candidates for inclusion in a portfolio. Most firms organize their analysts along traditional industry lines. One analyst might follow steel stocks or, perhaps in a smaller firm, all metal stocks. A second analyst might follow chemical stocks. But portfolio analysis calls for these analysts not only to estimate how a particular steel stock will move in relationship to another steel stock but also how a particular steel stock will move in relationship to a particular chemical stock or drug stock. There is no nonoverlapping organizational structure that allows such estimates to be directly produced.

[1]Whether the analyst's estimates contain information or whether one is better off estimating returns from an equilibrium model (such as those presented in Chapters 10, 13, and 14) is an open question. We have more to say about this later. However, the reader should note that portfolio selection models can help to answer this question.

The problem is made more complex by the number of estimates required. Most financial institutions follow between 150 and 250 stocks. To employ portfolio analysis, the institution needs estimates of between 150 and 250 expected returns and 150 and 250 variances. Let us see how many correlation coefficients it needs. If we let N stand for the number of stocks a firm follows, then it has to estimate ρ_{ij} for all pairs of securities i and j. The first index i can take on N values (one for each stock); the second can take on $(N - 1)$ values (remember $j \neq i$). This gives us $N(N - 1)$ correlation coefficients. However, because the correlation coefficient between stocks i and j is the same as that between stocks j and i, we have to estimate only $N(N - 1)/2$ correlations. The institution that follows between 150 and 250 stocks needs between 11,175 and 31,125 correlation coefficients. The sheer number of inputs is staggering.

It seems unlikely that analysts will be able to directly estimate correlation structures. Their ability to do so is severely limited by the nature of feasible organizational structures and the huge number of correlation coefficients that must be estimated. Recognition of this has motivated the search for the development of models to describe and predict the correlation structure between securities. In this chapter and in Chapter 8 we discuss some of these models and examine empirical tests of their performance.

The models developed for forecasting correlation structures fall into two categories: index models and averaging techniques. The most widely used technique assumes that the comovement between stocks is due to a single common influence or index. This model is appropriately called the *single-index model*. The single-index model is used not only in estimating the correlation matrix but also in efficient market tests (discussed later) and in equilibrium tests, where it is called a *return-generating process*. The rest of this chapter is devoted to a discussion of the properties of this model.

SINGLE-INDEX MODELS: AN OVERVIEW

Casual observation of stock prices reveals that when the market goes up (as measured by any of the widely available stock market indexes), most stocks tend to increase in price, and when the market goes down, most stocks tend to decrease in price. This suggests that one reason security returns might be correlated is because of a common response to market changes, and a useful measure of this correlation might be obtained by relating the return on a stock to the return on a stock market index. The return on a stock can be written as[2]

$$R_i = a_i + \beta_i R_m$$

where

- a_i is the component of security i's return that is independent of the market's performance—a random variable.
- R_m is the rate of return on the market index—a random variable.
- β_i is a constant that measures the expected change in R_i given a change in R_m.

This equation simply breaks the return on a stock into two components, that part due to the market and that part independent of the market. Variable β_i in the expression measures how sensitive a stock's return is to the return on the market. A β_i of 2 means that a stock's

[2]The return on the index is identical, in concept, to the return on a common stock. It is the return the investor would earn if she held a portfolio with a composition identical to the index. Thus, to compute this return, the dividends that would be received from holding the index should be calculated and combined with the price changes on the index.

return is expected to increase (decrease) by 2% when the market increases (decreases) by 1%. Similarly, a β of 0.5 indicates that a stock's return is expected to increase (decrease) by $\frac{1}{2}$ of 1% when the market increases (decreases) by 1%.[3]

The term a_i represents that component of return insensitive to (independent of) the return on the market. It is useful to break the term a_i into two components. Let α_i denote the expected value of a_i, and let e_i represent the random (uncertain) element of a_i. Then

$$a_i = \alpha_i + e_i$$

where e_i has an expected value of zero. The equation for the return on a stock can now be written as

$$R_i = \alpha_i + \beta_i R_m + e_i \tag{7.3}$$

Once again, note that both e_i and R_m are random variables. They each have a probability distribution and a mean and standard deviation. Let us denote their standard deviations by σ_{ei} and σ_m, respectively. Up to this point we have made no simplifying assumptions. We have written return as the sum of several components, but these components, when added together, must by definition be equal to total return.

It is convenient to have e_i uncorrelated with R_m. Formally, this means that

$$\text{cov}(e_i\ R_m) = E\left[(e_i - 0)\left(R_m - \overline{R}_m\right)\right] = 0$$

If e_i is uncorrelated with R_m, it implies that how well Equation (7.3) describes the return on any security is independent of what the return on the market happens to be. Estimates of α_i, β_i, and σ_{ei}^2 are often obtained from time series–regression analysis.[4] Regression analysis is one technique that guarantees that e_i and R_m will be uncorrelated, at least over the period to which the equation has been fit. All of the characteristics of single-index models described to this point are definitions or can be made to hold by construction. There is one further characteristic of single-index models: it holds only by assumption. This assumption is the characteristic of single-index models that differentiates them from other models used to describe the covariance structure.

The key assumption of the single-index model is that e_i is independent of e_j for all values of i and j, or, more formally, $E(e_i e_j) = 0$. This implies that the only reason stocks vary together, systematically, is because of a common comovement with the market. There are no effects beyond the market (e.g., industry effects) that account for comovement among securities. We will have more to say about this in our discussion of multi-index models in Chapter 8. However, at this time, note that, unlike the independence of e_i and R_m, there is nothing in the normal regression method used to estimate α_i, β_i, and σ_{ei}^2 that forces this to be true. It is a simplifying assumption that represents an approximation to reality. How well this model performs will depend, in part, on how good (or bad) this approximation is. Let us summarize the single-index model:

BASIC EQUATION

$R_i = \alpha_i + \beta_i R_m + e_i$ for all stocks $i = 1, ..., N$

[3]We are illustrating the single-index model with a stock market index. It is not necessary that the index used be a stock market index. The selection of the appropriate index is an empirical rather than a theoretical question. However, anticipating the results of future chapters, the results should be better when a broad-based market-weighted index is used, such as the S&P 500 index or the New York Stock Exchange index.

[4]This is discussed in more detail later in the chapter.

BY CONSTRUCTION

1. Mean of $e_i = E(e_i) = 0$	for all stocks $i = 1, ..., N$

BY ASSUMPTION

1. Index unrelated to unique return: $E[e_i(R_m - \bar{R}_m)] = 0$	for all stocks $i = 1, ..., N$
2. Securities related only through common response to market: $E(e_i e_j) = 0$	for all pairs of stocks $i = 1, ..., N$ and $j = 1, ..., N$ but $i \neq j$

BY DEFINITION

1. Variance of $e_i = E(e_i)^2 = \sigma_{ei}^2$	for all stocks $i = i, ..., N$
2. Variance of $R_m = E(R_m - \bar{R}_m)^2 = \sigma_m^2$	

In the subsequent section we derive the expected return, standard deviation, and covariance when the single-index model is used to represent the joint movement of securities. The results are

1. the mean return, $\bar{R}_i = \alpha_i + \beta_i \bar{R}_m$

2. the variance of a security's return, $\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{ei}^2$

3. the covariance of returns between securities i and j, $\sigma_{ij} = \beta_i \beta_j \sigma_m^2$

Note that the expected return has two components: a unique part α_i and a market-related part $\beta_i \bar{R}_m$. Likewise, a security's variance has the same two parts, unique risk σ_{ei}^2 and market-related risk $\beta_i^2 \sigma_m^2$. In contrast, the covariance depends only on market risk. This is what we meant earlier when we said that the single-index model implied that the only reason securities move together is a common response to market movements. In this section of the text we derive these results.

The expected return on a security is

$$E(R_i) = E[\alpha_i + \beta_i R_m + e_i]$$

Because the expected value of the sum of random variables is the sum of the expected values, we have

$$E(R_i) = E(\alpha_i) + E(\beta_i R_m) + E(e_i)$$

where α_i and β_i are constants, and by construction, the expected value of e_i is zero. Thus

$$E(R_i) = \alpha_i + \beta_i \bar{R}_m \qquad \textbf{Result 1}$$

The variance of the return on any security is

$$\sigma_i^2 = E(R_i - \bar{R}_i)^2$$

Substituting for R_i and $\bar{R}_i$ from the preceding expression yields

$$\sigma_i^2 = E[(\alpha_i + \beta_i R_m + e_i) - (\alpha_i + \beta_i \bar{R}_m)]^2$$

Rearranging and noting that the αs cancel yields

$$\sigma_i^2 = E[\beta_i (R_m - \bar{R}_m) + e_i]^2$$

Squaring the terms in the brackets yields

$$\sigma_i^2 = \beta_i^2 E(R_m - \bar{R}_m)^2 + 2\beta_i E[e_i (R_m - \bar{R}_m)] + E(e_i)^2$$

Recall that by assumption (or in some cases by construction), $E[e_i(R_m - \bar{R}_m)] = 0$. Thus

$$\sigma_i^2 = \beta_i^2 E\left(R_m - \bar{R}_m\right)^2 + E(e_i)^2$$
$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{ei}^2 \qquad \textbf{Result 2}$$

The covariance between any two securities can be written as

$$\sigma_{ij} = E\left[\left(R_i - \bar{R}_i\right)\left(R_j - \bar{R}_j\right)\right]$$

Substituting for R_i, $\bar{R}_i$, R_j, and $\bar{R}_j$ yields

$$\sigma_{ij} = E\left\{\left[\left(\alpha_i + \beta_i R_m + e_i\right) - \left(\alpha_i + \beta_i \bar{R}_m\right)\right]\right.$$
$$\left. \cdot \left[\left(\alpha_j + \beta_j R_m + e_j\right) - \left(\alpha_j + \beta_j \bar{R}_m\right)\right]\right\}$$

Simplifying by canceling the αs and combining the terms involving βs yields

$$\sigma_{ij} = E\left[\left(\beta_i\left(R_m - \bar{R}_m\right) + e_i\right)\left(\beta_j\left(R_m - \bar{R}_m\right) + e_j\right)\right]$$

Carrying out the multiplication,

$$\sigma_{ij} = \beta_i\beta_j E\left(R_m - \bar{R}_m\right)^2 + \beta_j E\left[e_i\left(R_m - \bar{R}_m\right)\right]$$
$$+ \beta_i E\left[e_j\left(R_m - \bar{R}_m\right)\right] + E\left(e_i e_j\right)$$

Because the last three terms are zero, by assumption,

$$\sigma_{ij} = \beta_i\beta_j\sigma_m^2 \qquad \textbf{Result 3}$$

These results can be illustrated with a simple example. Consider the returns on a stock and a market index shown in the first two columns of Table 7.1. These returns are what an investor might have observed over the prior five months. Now consider the values for the single-index model shown in the remaining columns of the table. Column 3 just reproduced column 1 and is the return on the security. In the next section of this chapter we will show you how to estimate α_i and β_i. For now assume that $\beta_i = 1.5$. This is what it would be equal to if we applied the first estimation technique described in the next section. Then, from result 1, $\alpha_i = 8 - 6 = 2$. Because the single-index model must hold as an equality, e_i (column 6) is just defined in each period as the value that makes the equality hold, or

$$e_i = R_i - \left(\alpha_i + \beta_i R_m\right)$$

For example, in the first period, the sum of α_i and $\beta_i R_m$ is 8. Because the return on the security in the first period is 10, e_i is +2.

Table 7.1 Decomposition of Returns for the Single-Index Model

Month	1 Return on Stock	2 Return on Market	3 R_i	=	4 α_i	+	5 $\beta_i R_m$	+	6 e_i
1	10	4	10	=	2	+	6	+	2
2	3	2	3	=	2	+	3	−	2
3	15	8	15	=	2	+	12	+	1
4	9	6	9	=	2	+	9	−	2
5	3	0	3	=	2	+	0	+	1
Total	40	20	40	=	10	+	30	+	0
Average	8	4	8	=	2	+	6	+	0
Variance	20.8	8	20.8	=	0	+	18	+	2.8

The reader should now understand where all the values of the single-index model come from, except β_i. Variable β_i divides return into market-related and unique return. When β_i is set equal to 1.5, the market return is independent of the residual return e_i. A lower value of e_i leaves some market return in e_i, and the covariance of e_i with the market is positive. A β_i greater than 1.5 removes too much market return and results in a negative covariance between e_i and the market. Thus the value of β_i is unique and is the value that exactly separates market from unique return, making the covariance between R_m and e_i zero. The reader can calculate the covariance between columns 2 and 6 in Table 7.1 and see that it is indeed zero.

Before leaving the simple example, let us apply the formulas presented earlier. The mean return on the security is

$$\bar{R}_i = 40/5 = 8$$

using the formula from the single-index model,

$$\bar{R}_i = \alpha_i + \beta_i \bar{R}_m = 2 + 1.5(4) = 8$$

The variance of security i is calculated from the formula derived for the single-index model:

$$\begin{aligned}\sigma_i^2 &= \beta_i^2 \sigma_m^2 + \sigma_{ei}^2 \\ &= (1.5)^2(8) + 2.8 \\ &= 20.8\end{aligned}$$

Calculating the variance of the security directly from column 1 of Table 7.1, we see that the answer is 20.8, identical to the answer produced by the preceding equation.

Having explained the simple example, we can turn to the calculation of the expected return and variance of any portfolio if the single-index model holds. The expected return on any portfolio is given by

$$\bar{R}_P = \sum_{i=1}^{N} X_i \bar{R}_i$$

Substituting for $\bar{R}_i$, we obtain

$$\bar{R}_P = \sum_{i=1}^{N} X_i \alpha_i + \sum_{i=1}^{N} X_i \beta_i \bar{R}_m \tag{7.4}$$

We know that the variance of a portfolio of stocks is given by

$$\sigma_P^2 = \sum_{i=1}^{N} X_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{\substack{j=1; \\ j \neq i}}^{N} X_i X_j \sigma_{ij}$$

Substituting in the results for σ_i^2 and σ_{ij}, we obtain

$$\sigma_P^2 = \sum_{i=1}^{N} X_i^2 \beta_i^2 \sigma_m^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \beta_i \beta_j \sigma_m^2 + \sum_{i=1}^{N} X_i^2 \sigma_{ei}^2 \tag{7.5}$$

There are many alternative ways of estimating the parameters of the single-index model. From Equations (7.4) and (7.7) it is clear that expected return and risk can be estimated for any portfolio if we have an estimate of α_i for each stock, an estimate of β_i for each stock, an estimate of σ_{ei}^2 for each stock, and, finally, an estimate of both the expected return ($\bar{R}_m$)

and the variance (σ_m^2) for the market. This is a total of $3N + 2$ estimates. For an institution following between 150 and 250 stocks, the single-index model requires between 452 and 752 estimates. Compare this with the 11,175–31,125 correlation estimates or 11,475–31,625 total estimates required when no simplifying structure is assumed. Furthermore, note that there is no requirement for direct estimates of the joint movement of securities, only estimates of the way each security moves with the market. A nonoverlapping organizational structure can produce all the required estimates.

The model can also be employed if analysts supply estimates of expected return for each stock, the variance of the return on each stock, the beta (β_i) for each stock, and the variance of the market return.[5] This is $3N + 1$ estimates. This alternative set of estimates has the advantage that they are in more familiar terms.

We have discussed means and variances before. The only new variable is beta, which is simply a measure of the sensitivity of a stock to market movements.

Before we discuss alternative ways of estimating betas, let us examine some of the characteristics of the single-index model.

CHARACTERISTICS OF THE SINGLE-INDEX MODEL

Define the beta on a portfolio β_P as a weighted average of the individual β_is on each stock in the portfolio, where the weights are the fraction of the portfolio invested in each stock. Then

$$\beta_P = \sum_{i=1}^{N} X_i \beta_i$$

Similarly, define the alpha on the portfolio α_P as

$$\alpha_P = \sum_{i=1}^{N} X_i \alpha_i$$

Then Equation (7.4) can be written as

$$\bar{R}_P = \alpha_P + \beta_P \bar{R}_m$$

If the portfolio P is taken to be the market portfolio (all stocks held in the same proportions as they were in constructing R_m), then the expected return on P must be $\bar{R}_m$. From the above equation the only values of β_P and α_P that guarantee $\bar{R}_P = \bar{R}_m$ for any choice of $\bar{R}_m$ are α_P equal to 0 and β_P equal to 1. Thus the beta on the market is 1 and stocks are thought of as being more or less risky than the market, according to whether their beta is larger or smaller than 1.

---

[5]The fact that these inputs are equivalent to those discussed earlier is easy to show. The expected returns can be used directly to estimate the expected return on a portfolio:

$$\bar{R}_P = \sum_{i=1}^{N} X_i \bar{R}_i$$

The estimates of the variance of return on a stock, the variance of the market, and the beta on each stock can be used to derive estimates of its residual risk by noting that

$$\sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{ei}^2$$

In addition, this structure is natural for those institutions that want analysts' estimates of means and variances and model estimates of correlations or covariances.

Let us look further into the risk of an individual security. Equation (7.5) is

$$\sigma_P^2 = \sum_{i=1}^{N} X_i^2 \beta_i^2 \sigma_m^2 + \sum_{i=1}^{N} X_i^2 \sigma_{ei}^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \beta_i \beta_j \sigma_m^2$$

In the double summation $i \neq j$, if $i = j$, then the terms would be $X_i X_i \beta_i^2 \sigma_m^2$. But these are exactly the terms in the first summation. Thus the variance on the portfolio can be written as

$$\sigma_P^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} X_i X_j \beta_i \beta_j \sigma_m^2 + \sum_{i=1}^{N} X_i^2 \sigma_{ei}^2$$

or by rearranging terms,

$$\sigma_P^2 = \left(\sum_{i=1}^{N} X_i \beta_i \right) \left(\sum_{j=1}^{N} X_j \beta_j \right) \sigma_m^2 + \sum_{i=1}^{N} X_i^2 \sigma_{ei}^2$$

Thus the risk of the investor's portfolio could be represented as

$$\sigma_P^2 = \beta_P^2 \sigma_m^2 + \sum_{i=1}^{N} X_i^2 \sigma_{ei}^2$$

Assume for a moment that an investor forms a portfolio by placing equal amounts of money into each of N stocks. The risk of this portfolio can be written as[6]

$$\sigma_P^2 = \beta_P^2 \sigma_m^2 + \frac{1}{N} \left(\sum_{i=1}^{N} \frac{1}{N} \sigma_{ei}^2 \right)$$

Look at the last term. This can be expressed as $1/N$ times the average residual risk in the portfolio. As the number of stocks in the portfolio increases, the importance of the average residual risk,

$$\sum_{i=1}^{N} \frac{\sigma_{ei}^2}{N}$$

diminishes drastically. In fact, as Table 7.2 shows, the residual risk falls so rapidly that most of it is effectively eliminated on even moderately sized portfolios.[7]

The risk that is not eliminated as we hold larger and larger portfolios is the risk associated with the term β_P. If we assume that residual risk approaches zero, the risk of the portfolio approaches

$$\sigma_P = \left[\beta_P^2 \sigma_m^2 \right]^{1/2} = \beta_P \sigma_m = \sigma_m \left[\sum_{i=1}^{N} X_i \beta_i \right]$$

Because σ_m is the same regardless of which stock we examine, the measure of the contribution of a security to the risk of a large portfolio is β_i.

---

[6]Examining the expression for the variance of portfolio P shows that the assumptions of the single-index model are inconsistent with $\sigma_P^2 = \sigma_m^2$. However, the approximation is very close. See Fama (1968) for a detailed discussion of this issue.

[7]To the extent that the single-index model is not a perfect description of reality and residuals from the market model are correlated across securities, residual risk does not fall this rapidly. However, for most portfolios, the amount of positive correlation present in the residuals is quite small, and residual risk declines rapidly as the number of securities in the portfolio increases.

Table 7.2 Residual Risk and Portfolio Size

Number of Securities	Residual Risk (Variance) Expressed as a Percentage of the Residual Risk Present in a One-Stock Portfolio with σ_{ei}^2 a Constant
1	100
2	50
3	33
4	25
5	20
10	10
20	5
100	1
1,000	0.1

The risk of an individual security is $\beta_i^2\sigma_m^2 + \sigma_{ei}^2$. Because the effect of σ_{ei}^2 on portfolio risk can be made to approach zero as the portfolio gets larger, it is common to refer to σ_{ei}^2 as diversifiable risk.[8] However, the effect of $\beta_i^2\sigma_m^2$ on portfolio risk does not diminish as N gets larger. Because σ_m^2 is a constant with respect to all securities, β_i is the measure of a security's nondiversifiable risk.[9] Because diversifiable risk can be eliminated by holding a large enough portfolio, β_i is often used as the measure of a security's risk.

ESTIMATING BETA

The use of the single-index model calls for estimates of the beta of each stock that is a potential candidate for inclusion in a portfolio. Analysts could be asked to provide subjective estimates of beta for a security or a portfolio. Conversely, estimates of future beta could be arrived at by estimating beta from past data and using this historical beta as an estimate of the future beta. There is evidence that historical betas provide useful information about future betas. Furthermore, some interesting forecasting techniques have been developed to increase the information that can be extracted from historical data. Because of this, even the firm that wishes to use analysts' subjective estimates of future betas should start with (supply analysts with) the best estimates of beta available from historical data. The analyst can then concentrate on the examination of influences that are expected to change beta in the future. In the rest of this chapter we examine some of the techniques that have been proposed for estimating beta. These techniques can be classified as measuring historical betas, correcting historical betas for the tendency of historical betas to be closer to the mean when estimated in a future period, and correcting historical estimates by incorporating fundamental firm data.

Estimating Historical Betas

In Equation (7.3) we represented the return on a stock as

$$R_i = \alpha_i + \beta_i R_m + e_i$$

This equation is expected to hold at each moment in time, although the value of α_i, β_i, or σ_{ei}^2 might differ over time. When looking at historical data, one cannot directly observe α_i,

[8]An alternative nomenclature calls this *nonmarket* or *unsystematic risk*.

[9]An alternative nomenclature calls this *market risk* or *systematic risk*.

β_i, or σ_{ei}^2. Rather, one observes the past returns on the security and the market. If α_i, β_i, and σ_{ei}^2 are assumed to be constant through time, then the same equation is expected to hold at each point in time. In this case, a straightforward procedure exists for estimating α_i, β_i, and σ_{ei}^2.

Notice that Equation (7.3) is an equation of a straight line. If σ_{ei}^2 were equal to zero, then we could estimate α_i and β_i with just two observations. However, the presence of the random variable e_i means that the actual return will form a scatter around the straight line. Figure 7.1 illustrates this pattern. The vertical axis is the return on security *i*, and the horizontal axis is the return on the market. Each point on the diagram is the return on stock *i* over a particular time interval, for example, one month (*t*) plotted against the return on the market for the same time interval. The actual observed returns lie on and around the true relationship (shown as a solid line). The greater σ_{ei}^2 is, the greater is the scatter around the line, and because we do not actually observe the line, the more uncertain we are about where it is. There are a number of ways of estimating where the line might be, given the observed scatter of points. Usually, we estimate the location of the line using regression analysis.

This procedure could be thought of as first plotting R_{it} versus R_{mt} to obtain a scatter of points such as that shown in Figure 7.1. Each point represents the return on a particular stock and the return on the market in one month. Additional points are obtained by plotting the two returns in successive months. The next step is to fit that straight line to the data that minimized the sum of the squared deviation from the line in the vertical (R_{it}) direction. The slope of this straight line would be our best estimate of beta over the period to which the line was fit, and the intercept would be our best estimate of alpha (α_i).[10]

More formally, to estimate the beta for a firm for the period from $t = 1$ to $t = 60$ via regression analysis, use

$$\beta_i = \frac{\sigma_{im}}{\sigma_m^2} = \frac{\sum_{t=1}^{60}\left[\left(R_{it} - \overline{R}_{it}\right)\left(R_{mt} - \overline{R}_{mt}\right)\right]}{\sum_{t=1}^{60}\left(R_{mt} - \overline{R}_{mt}\right)^2}$$

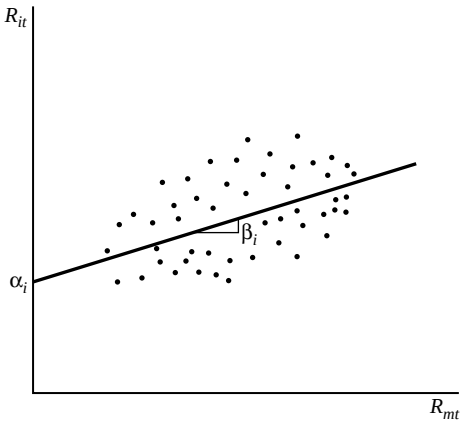


Figure 7.1 Plot of security return versus market return.

[10]If R_{it} and R_{mt} come from a bivariate normal distribution, the unbiased and most efficient estimates of α_i and β_i are those that come from regressing R_{it} against R_{mt}, the procedure described earlier.

and to estimate alpha, use[11]

$$\alpha_i = \overline{R}_{it} - \beta_i \overline{R}_{mt}$$

To learn how this works on a simple example, let us return to Table 7.1. We used the data in Table 7.1 to show how beta interacted with returns. But now assume that all you observed were columns 1 and 2 or the return on the stock and the return on the market. To estimate beta, we need to estimate the covariance between the stock and the market. The average return on the stock was 40/5 = 8, whereas on the market it was 20/5 = 4. The beta value for the stock is the covariance of the stock with the market divided by the variance of the market, or

$$\beta_i = \frac{\left[\sum_{t=1}^{5}\left(R_i - \overline{R}_i\right)\left(R_m - \overline{R}_m\right)\right] \Big/ 5}{\sum_{t=1}^{5}\left(R_m - \overline{R}_m\right)^2 \Big/ 5}$$

The covariance is found as follows:

Month	Stock Return Minus Mean		Market Return Minus Mean		Value
1	(10 − 8)	×	(4 − 4)	=	0
2	(3 − 8)	×	(2 − 4)	=	10
3	(15 − 8)	×	(8 − 4)	=	28
4	(9 − 8)	×	(6 − 4)	=	2
5	(3 − 8)	×	(0 − 4)	=	20
				Total	60

The covariance is 60/5 = 12. The variance of the market return is the average of the sum of squared deviation from the mean:

$$\sigma_m^2 = \left[(4-4)^2 + (2-4)^2 + (8-4)^2 + (6-4)^2 + (0-4)^2\right] \Big/ 5 = 8$$

Thus beta = 12/8 = 1.5. This value of beta is identical to the number used in constructing Table 7.1.

Alpha can be computed by taking the difference between the average security return and beta times the average return on the market:

$$\alpha_i = 8 - (1.5)(4) = 2$$

[11]Two other statistics of interest can be produced by this analysis. First, the size of σ_{ei}^2 over the estimation period can be found by looking at the variance of the deviations of the actual return from that predicted by the model:

$$\sigma_{ei}^2 = \frac{1}{60}\sum_{t=1}^{60}\left[R_{it} - \left(\alpha_i + \beta_i R_{mt}\right)\right]^2$$

Remember that in performing regression analysis, one often computes a coefficient of determination. The coefficient of determination is a measure of association between two variables. In this case, it would measure how much of the variation in the return on the individual stock is associated with variation in the return on the market. The coefficient of determination is simply the correlation coefficient squared, and the correlation coefficient is equal to

$$\rho_{im} = \frac{\sigma_{im}}{\sigma_i \sigma_m} = \frac{\beta_i \sigma_m^2}{\sigma_i \sigma_m} = \beta_i \frac{\sigma_m}{\sigma_i}$$

The values of α_i and β_i produced by regression analysis are estimates of the true α_i and β_i that exist for a stock. The estimates are subject to error. As such, the estimate of α_i and β_i may not be equal to the true α_i and β_i that existed in the period.[12] Furthermore, the process is complicated by the fact that α_i and β_i are not perfectly stationary over time. We would expect changes as the fundamental characteristics of the firm change. For example, β_i as a risk measure should be related to the capital structure of the firm and thus should change as the capital structure changes.

Despite error in measuring the true β_i and the possibility of real shifts in β_i over time, the most straightforward way to forecast β_i for a future period is to use an estimate of β_i obtained via regression analysis from a past period. Let us take a look at how well this works.

Accuracy of Historical Betas

The first logical step in looking at betas is to see how much association there is between the betas in one period and the betas in an adjacent period. Both Blume (1970) and Levy (1971) have done extensive testing of the relationship between betas over time. Let us look at some representative results from Blume's (1970) study. Blume computed betas using time series regressions on monthly data for nonoverlapping seven-year periods. He generated betas on single-stock portfolios, 2-stock portfolios, 4-stock portfolios, and so forth, up to 50-stock portfolios, and for each size portfolio, he examined how highly correlated the betas from one period were with the betas for a second period. Table 7.3 presents a typical result showing how highly correlated the betas are for the period 7/54–6/61 and 7/61–6/68.

It is apparent from this table that, while betas on very large portfolios contain a great deal of information about future betas on these portfolios, betas on individual securities contain much less information about the future betas on securities. Why might observed betas in one period differ from betas in a second period? One reason is that the risk (beta) of the security or portfolio might change. A second reason is that the beta in each period is measured with a random error, and the larger the random error, the less predictive power betas from one period will have for betas in the next period.

Changes in security betas will differ from security to security. Some will go up, some will go down. These changes will tend to cancel out in a portfolio, and we observe less change in the actual beta on portfolios than on securities.

Table 7.3 Association of Betas over Time

Number of Securities in the Portfolio	Correlation Coefficient	Coefficient of Determination
1	0.60	0.36
2	0.73	0.53
4	0.84	0.71
7	0.88	0.77
10	0.92	0.85
20	0.97	0.95
35	0.97	0.95
50	0.98	0.96

[12]In fact, the analysis will produce an estimate of the standard error in both α_i and β_i. This can be used to make interval estimates of future alphas and betas under the assumption of stationarity.

Likewise, one would expect that the errors in estimating beta for individual securities would tend to cancel out when securities are combined, and therefore, there would be less error in measuring a portfolio's beta.[13] Because portfolio betas are measured with less error, and because betas on portfolios change less than betas on securities, historical betas on portfolios are better predictors of future betas than are historical betas on securities.

Adjusting Historical Estimates

Can we further improve the predictive ability of betas on securities and portfolios? To aid in answering this question, let us examine a simple hypothetical distribution of betas. Assume the true betas on all stocks are really 1. If we estimate betas for all stocks, some of our estimated betas will be 1, but some will be above or below 1 owing to sampling error in the estimate. Estimated betas above 1 would be above 1 simply because of positive sampling errors. Estimated betas below 1 would be below 1 because of negative sampling errors. Furthermore, because there is no reason to suspect that a positive sampling error for a stock will be followed by a positive sampling error for the same stock, we would find that historical beta did a worse job of predicting future beta than did a beta of 1 for all stocks. Now, assume we have different betas for different stocks. The beta we calculate for any stock will be, in part, a function of the true underlying beta and, in part, a function of sampling error. If we compute a very high estimate of beta for a stock, we have an increased probability that we have a positive sampling error, whereas if we compute a very low estimate of beta, we have an increased chance that we have a negative sampling error. If this scenario is correct, we should find that betas, on the average, tend to converge to 1 in successive time periods. Estimated betas that are a lot larger than 1 should tend to be followed by estimated betas that are closer to 1 (lower), and estimated betas below 1 should tend to be followed by higher betas. Evidence that this does, in fact, happen has been presented by Blume (1975) and Levy (1971). Blume's results are reproduced in Table 7.4. The reader should examine the table and confirm the tendency of betas in the forecast period to be closer to 1 than the estimates of these betas obtained from historical data.[14]

---

[13]Assuming that the relationship between R_{it} and R_{mt} is described by a stationary bivariate normal distribution, then the standard error in the measurement of beta for a security is given by

$$\sigma_{\beta i} = \sigma_{ei}/\sigma_m$$

The standard error for the β on a portfolio is given by

$$\sigma_{\beta P} = \sigma_{ep}/\sigma_m$$

where

$$\sigma_{ep}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(e_{pt}\right)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i=1}^{N} X_i e_{it}\right)^2$$

where N is the number of securities in the portfolio and T is the number of time periods.

To the extent that the residuals for different stocks are not perfectly correlated, averaging them across stocks will lower the value of the residuals and, hence, the value of σ_{ep}^2 on the portfolio. In particular, if the assumptions of the single-index model are met, and if stocks are held in equal proportions, the standard error of the beta on the portfolio would equal the average standard error on all stocks times the reciprocal of the number of stocks in the portfolio.

[14]Throughout this section, when we speak of betas, we are referring to estimates of betas.

Table 7.4 Betas on Ranked Portfolios for Two Successive Periods

Portfolio	7/54–6/61	7/61–6/68
1	0.393	0.620
2	0.612	0.707
3	0.810	0.861
4	0.987	0.914
5	1.138	0.995
6	1.337	1.169

Source: Blume, Marchell. "On the Assessment of Risk," *Journal of Finance*, **VI**, No. 1 (March 1971) p. 8.

Measuring the Tendency of Betas to Regress toward 1—Blume's Technique

Because betas in the forecast period tend to be closer to 1 than the estimate obtained from historical data, the next obvious step is to try to modify past betas to capture this tendency. Blume (1975) was the first to propose a scheme for doing so. He corrected past betas by directly measuring this adjustment toward 1 and assuming that the adjustment in one period is a good estimate of the adjustment in the next.

Let us see how this could work. We could calculate the betas for all stocks for the period 1948–1954. We could then calculate the betas for these same stocks for the period 1955–1961. We could then regress the betas for the later period against the betas for the earlier period, as shown in Figure 7.2. Note that each observation is the beta on the same stock for the period 1948–1954 and 1955–1961. Following this procedure, we would obtain a line that measures the tendency of the forecasted betas to be closer to 1 than the estimates from historical data. When Blume did this for the period mentioned, he obtained

$$\beta_{i2} = 0.343 + 0.677\beta_{i1}$$

where β_{i2} stands for the beta on stock i in the later period (1955–1961) and β_{i1} stands for the beta for stock i for the earlier period (1948–1954). The relationship implies that the beta in the later period is 0.343 + 0.677 times the beta in the earlier period. Assume we wish to

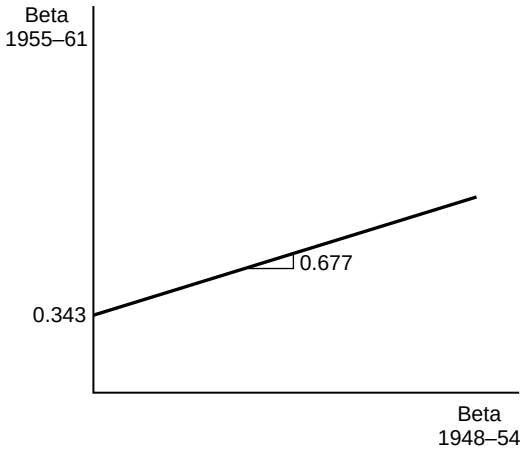


Figure 7.2 Plot of beta in two adjacent periods.

forecast the beta for any stock for the period 1962–1968. We then compute (via regression analysis) its beta for the years 1955–1961. To determine how this beta should be modified, we substitute it for β_{i1} in the equation. We then compute β_{i2} from the foregoing equation and use it as our forecast.

Notice the effect of this on the beta for any stock. If β_{i1} were 2.0, then our forecast would be $0.343 + 0.677(2) = 1.697$ rather than 2.0. If β_{i1} were 0.5, our forecast would be $0.343 + 0.677(0.5) = 0.682$ rather than 0.5. The equation lowers high values of beta and raises low values. One more characteristic of this equation should be noted: it modifies the average level of betas for the population of stocks. Because it measures the relationship between betas over two periods, if the average beta increased over these two periods, it assumes that average betas will increase over the next period. Unless there is reason to suspect a continuous drift in beta, this will be an undesirable property. If there is no reason to expect this trend in the average beta to continue, then the estimates can be improved by adjusting the forecasted betas so that their mean is the same as the historical mean.

To make this point more concrete, let us examine an example. Assume that in estimating the equation, Blume found that the average beta in 1948–1954 was 1 and the average beta in 1955–1961 was 1.02. These numbers are consistent with his results, though there are other sets of numbers that would also be consistent with his results. Now, to determine what the average forecasted beta should be for the period 1962–1968, we simply substitute 1.02 into the right-hand side of the estimating equation. The answer is 1.033. As discussed earlier, Blume's technique results in a continued extrapolation of the upward trend in betas observed in the earlier periods.

If there is no reason to believe that the next period's average beta will be more than this period's, then the forecasts should be improved by adjusting the forecast beta to have the same mean as the historical mean. This involves subtracting a constant from all betas after adjusting them toward their mean. In our example, this is achieved by subtracting 1.033 from each forecast of beta and adding 1.02.

Measuring the Tendency of Betas to Regress toward 1—Vasicek's Technique

Recall that the actual beta in the forecast period tends to be closer to the average beta than is the estimate obtained from historical data. A straightforward way to adjust for this tendency is simply to adjust each beta toward the average beta. For example, taking one-half of the historical beta and adding it to one-half of the average beta moves each historical beta halfway toward the average. This technique is widely used.[15]

It would be desirable not to adjust all stocks the same amount toward the average but rather to have the adjustment depend on the size of the uncertainty (sampling error) about beta. The larger the sampling error, the greater the chance of large differences from the average being due to sampling error, and the greater the adjustment. Vasicek (1973) has suggested the following scheme that incorporates these properties: if we let $\bar{\beta}_1$ equal the average beta across the sample of stocks in the historical period, then the Vasicek procedure involves taking a weighted average of $\bar{\beta}_1$ and the historical beta for security *i*. Let $\sigma^2_{\bar{\beta}1}$ stand for the variance of the distribution of the historical estimates of beta over the sample of stocks. This is a measure of the variation of beta across the sample of stocks under consideration. Let $\sigma^2{}_{\beta i1}$ stand for the square of the standard error of the estimate of beta for security *i*

[15]For example, Merrill Lynch has used a simple weighting technique like this to adjust its betas.

measured in time period 1. This is a measure of the uncertainty associated with the measurement of the individual securities beta. Vasicek (1973) suggested weights of

$$\frac{\sigma^2_{\bar{\beta}1}}{\sigma^2_{\bar{\beta}1} + \sigma^2_{\beta i1}} \text{ for } \beta_{i1} \qquad \text{and} \qquad \frac{\sigma^2_{\beta i1}}{\sigma^2_{\bar{\beta}1} + \sigma^2_{\beta i1}} \text{ for } \bar{\beta}_1$$

Note that these weights add up to 1 and that the more the uncertainty about either estimate of beta, the lower the weight that is placed on it. The forecast of beta for security *i* is

$$\beta_{i2} = \frac{\sigma^2_{\beta i1}}{\sigma^2_{\bar{\beta}1} + \sigma^2_{\beta i1}} \bar{\beta}_1 + \frac{\sigma^2_{\bar{\beta}1}}{\sigma^2_{\bar{\beta}1} + \sigma^2_{\beta i1}} \beta_{i1}$$

This weighting procedure adjusts observations with large standard errors further toward the mean than it adjusts observations with small standard errors. As Vasicek has shown, this is a Bayesian estimation technique.[16]

Although the Bayesian technique does not forecast a trend in betas as does the Blume technique, it suffers from its own potential source of bias. In the Bayesian technique, the weight placed on a stock's beta, relative to the weight on the average beta in the sample, is inversely related to the stock's standard error of beta. High-beta stocks have larger standard errors associated with their betas than do low-beta stocks. This means that high-beta stocks will have their betas lowered by a bigger percentage of the distance from the average beta for the sample than low-beta stocks will have their betas raised. Hence the estimate of the average future beta will tend to be lower than the average beta in the sample of stocks over which betas are estimated.

Unless there is reason to believe that betas will continually decrease, the estimate of beta can be further improved by adjusting all betas upward so that they have the same mean as they had in the historical period.

Accuracy of Adjusted Beta

Let us examine how well the Blume and the Bayesian adjustment techniques worked as forecasters, compared to unadjusted betas. Klemkosky and Martin (1975) tested the ability of these techniques to forecast over three five-year periods for both 1-stock and 10-stock portfolios. As would be suspected, in all cases both the Blume and Bayesian adjustment techniques led to more accurate forecasts of future betas than did the unadjusted betas. The average squared error in forecasting beta was often cut in half when one of the adjustment techniques was used. Klemkosky and Martin used an interesting decomposition technique to search for the source of the forecast error. Specifically, the source of error was decomposed into that part of the error due to a misestimate of the average level of beta, that part due to the tendency to overestimate high betas and underestimate low betas, and that part that is unexplained by either of the first two influences. As might be expected, when the Blume and Bayesian techniques were compared with the unadjusted betas, almost all of the decrease in error came from the reductions in the tendency to overestimate high betas and underestimate low betas. This is not surprising because this is exactly what the two techniques were designed to achieve. Klemkosky and Martin found that the Bayesian technique had a slight tendency to outperform the Blume technique.

[16]The reader should note that this is just one of an infinite number of ways of forming prior distributions. For example, priors could have been set equal to 1 (the average for all stocks market weighted) or to an average beta for the industry to which the stock belongs.

However, the differences were small, and the ordering of the techniques varied across different periods of time.

Most of the literature dealing with betas has evaluated beta adjustment techniques by their ability to better forecast betas. However, there is another, and perhaps more important, criterion by which the performance of alternative betas can be judged. At the beginning of this chapter we discussed the fact that the necessary inputs to portfolio analysis were expected returns, variances, and correlations. We believe that analysts can be asked to provide estimates of expected returns and variances but that correlations will probably continue to be generated from some sort of historical model.[17] One way betas can be used is to generate estimates of the correlation between securities. The correlations between stocks (given the assumptions of the single-index model) can be expressed as a function of beta:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\beta_i \beta_j \sigma_m^2}{\sigma_i \sigma_j}$$

Another way to test the usefulness of betas, as well as the performance of alternative forecasts of betas, is to see how well betas forecast the correlation structure between securities.

Betas as Forecasters of Correlation Coefficients

Elton, Gruber, and Urich (1978) have compared the ability of the following models to forecast the correlation structure between securities:

1. the historical correlation matrix itself
2. forecasts of the correlation matrix prepared by estimating betas from the prior historical period
3. forecasts of the correlation matrix prepared by estimating betas from the prior two periods and updating via the Blume technique
4. forecasts prepared as in the third model but where the updating is done via the Vasicek Bayesian technique

One of the most striking results of the study was that the historical correlation matrix itself was the poorest of all techniques. In most cases it was outperformed by all of the beta forecasting techniques at a statistically significant level. This indicates that a large part of the observed correlation structure between securities, not captured by the single-index model, represents random noise with respect to forecasting. The point to note is that the single-index model, developed to simplify the inputs to portfolio analysis and thought to lose information because of the simplification involved, actually does a better job of forecasting than the full set of historical data.

The comparison of the three beta techniques is more ambiguous. In each of two five-year samples tested, the Blume adjustment technique outperformed both the unadjusted betas and the betas adjusted via the Bayesian technique. The difference in the techniques was statistically significant. However, the Bayesian adjustment technique performed better than the unadjusted beta in one period and worse in a second. In both cases, the results were statistically significant. This calls for some further analysis. The performance of any forecasting technique is, in part, a function of its forecast of the average correlation between all stocks and, in part, a function of its forecast of previous differences from the mean.

[17] It is possible that analysts will be used to subjectively modify historical estimates of beta to improve their accuracy. Several firms currently use analysts' modified estimates of beta.

We might stop for a moment and see why each of the beta techniques might produce forecasts of the average correlation coefficient between all stocks that are different from the average correlation coefficient in the data to which the technique is fitted.

Let us start with the unadjusted betas. This model assumes that the only correlation between stocks is one due to common correlation with the market. It ignores all other sources of correlation such as industry effects. To the extent that there are other sources of correlation that are, on the whole, positive, this technique will underestimate the average correlation coefficient in the data to which it is fitted. This is exactly what Elton, Gruber, and Urich (1978) showed happened in both periods over which the model was fitted.

The Blume technique suffers from the same bias, but it has two additional sources of bias. One is that the Blume technique adjusts all betas toward 1. This tends to raise the average correlation coefficient estimated from the Blume technique. The correlation coefficient is the product of two betas. To the extent that betas are reduced to 1 symmetrically (with no change in mean), the cross-products between them will tend to be larger. For example, the product of 1.1 and 0.9 is larger than the product of 1.2 and 0.8. There is another source of potential problems in the Blume technique. Remember that the Blume technique adjusts the betas in period 2 for the changes in betas between period 1 and period 2. If the average change in beta between periods 1 and 2 is positive (negative), the Blume technique will adjust the average beta for period 2 up (down).[18] In the Elton, Gruber, and Urich study, there was an upward drift in betas over the period studied, and this, combined with the tendency of the Blume technique to shrink all betas toward 1, resulted in forecasts of an average correlation coefficient well above the average correlation coefficient for the sample to which the model was fitted.

The Bayesian adjustment to betas, like the Blume adjustment, has some upward forecast bias because of its tendency to shrink betas toward 1, but it does not continue to project a trend in betas and, hence, correlation coefficients, as the Blume technique does. However, as pointed out earlier, it has a new source of bias—one that tends to pull betas and correlation coefficients in a downward direction. This occurs because high-beta stocks are adjusted more toward the mean than low-beta stocks.

Short of empirical tests, it is difficult to say whether, given any set of data, the alternative sources of bias, which work in different directions, will increase or decrease the forecast accuracy of the result. We do know that unless there are predictable trends in average correlation coefficient, the effect of these biases on forecast accuracy will be random from period to period. This source of randomness can be eliminated. One way to do it is to force the average correlation coefficient, estimated by each technique, to be the same and to be equal to the average correlation coefficient that existed in the period over which the model was fitted. If correlation coefficients do not have stable trends, this will be an efficient forecast procedure. It uses only available data and is also easy to do.

When the adjustments were made, the Bayesian adjustment produced the most accurate forecasts of the future correlation matrix. Its difference from the Blume technique, the unadjusted beta, and the historical matrix was statistically significant in all periods tested. The second-ranked technique varied through time with the Blume adjustment, outperforming the unadjusted beta in one period and being outperformed by the unadjusted beta in one period.[19]

---

[18]This would be a desirable property if trends in average correlation coefficients were expected to persist over time, but we see no reason to expect them to do so.

[19]In addition, tests were made that forced the average correlation coefficient from each technique to be the same and equal to the average correlation coefficient that occurred in the forecast period. This is equivalent to perfect foresight with respect to the average correlation coefficient. The rankings were the same as those discussed earlier when this was done.

The forecasts from the three beta techniques were compared with the forecasts from a fourth beta estimate, beta equals 1 for all stocks, as well as with the historical correlation matrix, as a forecast of the future. The mean forecast was adjusted to be the same for all techniques. The performance of the historical correlation matrix and the beta-equals-1 model was inferior to the performance of all other models at a statistically significant level.

Let us stop a minute and review the work on estimating betas. There are two reasons for estimating betas: the first is in order to forecast future betas; the second is to generate correlation coefficients as input to the portfolio problem. Empirical evidence strongly suggests that to forecast future betas, one should use either the Bayesian adjustment or the Blume adjustment rather than unadjusted betas. The evidence on the choice between the Blume and the Bayesian adjustment is mixed, but the Bayesian adjustment seems to work slightly better.

If the goal is estimating the future correlation matrix as an input to the portfolio problems, things get more complex. Unadjusted betas and adjusted betas, both by the Bayesian and the Blume techniques, all contain potential bias as forecasters of future correlation matrices.[20] The forecasts from all of these techniques can be examined directly, or the forecasts can be adjusted to remove bias in the forecast of the average correlation coefficient. The first fact to note is that each of these three estimates of beta outperforms the historical correlation matrix as a forecast of the future correlation matrix. Second, note that when compared to a beta of 1, all produce better forecasts. The ranking among these three techniques is a function of whether we make the adjustment to the average forecast. Because we believe it is appropriate to do so, we find that the Bayesian adjustment technique performs best. In Chapter 8 we discuss forecasting future correlation coefficients using a combination of past betas and other forecasts derived from historical data.

Recently, attempts have been made to incorporate more data than past return information into the forecasts of betas. We now take a brief look at some of the work that has been done in this area.

Fundamental Betas

Beta is a risk measure that arises from the relationship between the return on a stock and the return on the market. However, we know that the risk of a firm should be determined by some combination of the firm's fundamentals and the market characteristics of the firm's stock. If these relationships could be determined, they would help us to better understand and forecast betas.

One of the earliest attempts to relate the beta of a stock to fundamental firm variables was performed by Beaver, Kettler, and Scholes (1970). They examined the relationship between seven firm variables and the beta on a company's stock. The seven variables they used were:

1. dividend payout (dividends divided by earnings)
2. asset growth (annual change in total assets)
3. leverage (senior securities divided by total assets)
4. liquidity (current assets divided by current liabilities)
5. asset size (total assets)
6. earning variability (standard deviation of the earnings price ratio)
7. accounting beta (the beta that arises from a time series regression of the earnings of the firm against average earnings for the economy, often called the earnings beta)

[20]As discussed earlier, a smaller set of potential biases is present when betas are estimated.

An examination of these variables would lead us to expect a negative relationship between dividend payout and beta under one of two arguments:

1. Because management is more reluctant to cut dividends than raise them, high payout is indicative of confidence on the part of management concerning the level of future earnings.
2. Dividend payments are less risky than capital gains; hence, the company that pays out more of its earnings in dividends is less risky.

Growth is usually thought of as positively associated with beta. High-growth firms are thought of as more risky than low-growth firms.

Leverage tends to increase the volatility of the earnings stream, hence to increase risk and beta.

A firm with high liquidity is thought to be less risky than one with low liquidity, and hence liquidity should be negatively related to market beta.

Large firms are often thought to be less risky than small firms, if for no other reason than that they have better access to the capital markets. Hence they should have lower betas.

Finally, the more variable a company's earning stream and the more highly correlated it is with the market, the higher its beta should be.

Table 7.5 reports some of the results from the Beaver, Kettler, and Scholes (1970) study. Note all variables had the sign that we expected.

The next logical step in developing fundamental betas is to incorporate the effects of relevant fundamental variables simultaneously into the analysis. This is usually done by relating beta to several fundamental variables via multiple regression analysis.

An equation of the following form is estimated:

$$\beta_i = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_N X_N + e_i \tag{7.6}$$

where each X_i is one of the N variables hypothesized as affecting beta. Several studies have been performed that link beta to a set of fundamental variables, such as that studied by Beaver, Kettler, and Scholes (1970).[21] The list of variables that has been studied and linked

Table 7.5 Correlation between Accounting Measures of Risk and Market Beta

	Period 1 1947–1956		Period 2 1957–1965	
Variable	One-Stock Portfolio	Five-Stock Portfolio	One-Stock Portfolio	Five-Stock Portfolio
Payout	−0.50	−0.77	−0.24	−0.45
Growth	0.23	0.51	0.03	0.07
Leverage	0.23	0.45	0.25	0.56
Liquidity	−0.13	−0.44	−0.01	−0.01
Size	−0.07	−0.13	−0.16	−0.30
Earnings variability	0.58	0.77	0.36	0.62
Earnings beta	0.39	0.67	0.23	0.46

[21]For examples of the use of fundamental data to estimate betas, see Cohen, Schwartz, and Whitecomb (1978), Francis (1975), Hawawini and Vora (1980), Blume (1975), and Hill and Stone (1980). The ability of fundamental data to aid in the prediction of future betas has been mixed. Some studies find large improvements in forecasting ability, while others do not.

to betas is too long to review here. For example, Thompson (1978) reviews 43 variables, while Rosenberg and Marathe (1975) review 101. Rather than discuss the long list of variables that has been used to generate fundamental betas, let us review the relative strengths and weaknesses of fundamental and historical betas as well as one system, proposed by Barr Rosenberg (1976, 1975, 1973), that has been put forth to combine both types of betas.

The advantage of betas based on historical return data is that they measure the response of each stock to market movements. The disadvantage of this type of beta is that it reflects changes in the size or importance of company characteristics only after a long period of time has passed. For example, assume a company increased its debt-to-equity ratio. We would expect its beta to increase. However, if we are using 60 months of return data to estimate beta, one month after the company increased its debt-to-equity ratio, only one of the 60 data points will reflect the new information. Thus the change in debt-to-equity ratio would have only a very minor impact on the beta computed from historical return data. Similarly, one full year after the event, only 12 of the 60 data points used to measure beta will reflect the event.

Conversely, fundamental betas respond quickly to a change in the companies' characteristics because they are computed directly from these characteristics. However, the weakness of fundamental betas is that they are computed under the assumption that the responsiveness of all betas to an underlying fundamental variable is the same. For example, they assume that the beta for IBM will change in exactly the same way with a given change in its debt-to-equity ratio as will the beta of General Motors (GM).[22]

By combining the techniques of historical betas and fundamental betas into one system, Barr Rosenberg hopes to gain the advantages of each without being subject to the disadvantages of either. In addition, because Rosenberg and McKibben (1973) found that there were persistent differences between the betas of different industries, Rosenberg and Marathe (1975) introduced a set of industry dummy variables into the analysis to capture these differences. Rosenberg's system can be described as follows:[23]

$$\beta_i = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_7x_7 + a_8x_8 + \cdots + a_{46}x_{46} \tag{7.7}$$

where

- x_1 represents 14 descriptions of market variability. These 14 descriptions include historical values of beta as well as other market characteristics of the stock such as share trading, volume, and stock price range.
- x_2 represents seven descriptors of earnings variability. These descriptors include measures of earnings variability, earnings betas, and measures of the unpredictability of earnings such as the amount of extraordinary earnings reported.
- x_3 represents eight descriptors of unsuccess and low valuation. These descriptors include growth in earnings, the ratio of book value to stock price, relative strength, and other indicators of perceived success.
- x_4 represents nine descriptors of immaturity and smallness. These descriptors include total assets, market share, and other indicators of size and age.
- x_5 represents nine descriptors of growth orientation. These descriptors include dividend yield, earnings price ratios, and other measures of historical and perceived growth.

[22]Each of the regression coefficients of Equation (7.6) (e.g., a_1) has only one value for all firms. This means that a change of 1 unit in X_1 will change the beta of every firm by a_1 units.

[23]Rosenberg changes the variables in his system over time. This description is based on his system as it existed at a point in time as described in Rosenberg and Marathe (1975).

x_6 represents nine descriptors of financial risk. These include measures of leverage, interest coverage, and liquidity.

x_7 represents six descriptors of firm characteristics. These include indicators of stock listings and broad types of business.

x_8 through x_{46} are industry dummy variables. These variables allow the fact that different industries tend to have different betas, all other variables held constant, to be taken into account.

While conceptually, the Rosenberg technique is easy to grasp, the multitude of variables (101) makes it difficult to grasp the meaning of the parameterized model. The reason for moving to this complex model is to improve forecasting ability. Rosenberg and Marathe's (1975) initial testing indicates that the model involving both fundamental data and historical betas leads to better estimates of future betas than the use of either type of estimate in isolation.

Before ending this chapter, we should mention one more type of model that is beginning to attract attention. The Rosenberg system quickly reflects changes in beta that have occurred because it uses data that reflect present conditions (fundamental firm variables) to modify historical betas as forecasts of the future. A more ideal system would employ forecasts of future fundamental firm variables to modify historical estimates of beta—in other words, substitute estimates of future values on the right-hand side of Equation (7.7) rather than concurrent values. Now it seems unlikely that analysts can do this for the 101 variables used in Rosenberg's system. However, simpler systems employing a much smaller number of variables are being used in this way.

THE MARKET MODEL

Although the single-index model was developed to aid in portfolio management, a less restrictive form—known as the market model—has found increased usage in finance. The market model is identical to the single-index model, except that the assumption that $\text{cov}(e_i e_j) = 0$ is not made.[24]

The model starts with the simpler linear relationship of returns and the market,

$$R_i = \alpha_i + \beta_i R_m + e_i$$

and produces an expected value for any stock of

$$\overline{R}_i = \alpha_i + \beta_i \overline{R}_m$$

Because it does not make the assumption that all covariances among stocks are due to a common covariance with the market, however, it does not lead to the simple expressions of portfolio risk that arise under the single-index model.

We will meet the market model again as we progress through this book. It is used extensively in Chapter 17 on the efficient market. The point to keep in mind is that the discussion of estimating beta is equally as applicable whether we are talking about the market model or the single-index model.

[24]Actually, although the single-index model can be defined in terms of any influence (e.g., the rate of return on liverwurst), we usually think of the index as the rate of return on some market portfolio. The market model is always defined in terms of a market portfolio.

AN EXAMPLE

A manager of a large pension fund will often utilize several domestic stock managers. The pension fund sponsor (manager) can view the asset allocation problem as equivalent to selecting among various stock mutual funds. The data for the portfolios being considered by a large pension fund are as follows:

NAME	α_i	β_i	σ_{ei}^2
1. Small stock	6	1.4	65
2. Value	4	0.8	20
3. Growth	4.5	1.3	45
4. Large capitalization	0.8	0.90	24
5. Special situation	0.2	1.1	45

The alphas, betas, and residual risks were initially computed by running a regression of each fund's return on the return of the S&P index using five years of monthly returns. These estimates were then modified by the plan sponsor to reflect their beliefs. Management projected that the S&P index at this point had an expected return of 12.5% and an estimated standard deviation of return of 14.9%. The expected returns, standard deviations of return, and covariance using the single-index model are

$$\bar{R}_1 = 6 + 1.4(12.5) = 23.5$$
$$\bar{R}_2 = 4 + 0.8(12.5) = 14$$
$$\bar{R}_3 = 4.5 + 1.3(12.5) = 20.75$$
$$\bar{R}_4 = 0.8 + 0.9(12.5) = 12.05$$
$$\bar{R}_5 = 0.2 + 1.1(12.5) = 13.95$$
$$\sigma_1 = \left[(1.4)^2(14.9)^2 + 65\right]^{1/2} = 22.36$$
$$\sigma_2 = \left[(0.8)^2(14.9)^2 + 20\right]^{1/2} = 12.73$$
$$\sigma_3 = \left[(1.3)^2(14.9)^2 + 45\right]^{1/2} = 20.5$$
$$\sigma_4 = \left[(.90)^2(14.9)^2 + 24\right]^{1/2} = 14.28$$
$$\sigma_5 = \left[(1.1)^2(14.9)^2 + 45\right]^{1/2} = 17.71$$
$$\sigma_{12} = (1.4)(0.8)(14.9)^2 = 249$$
$$\sigma_{13} = (1.4)(1.3)(14.9)^2 = 404$$
$$\sigma_{14} = (1.4)(0.9)(14.9)^2 = 280$$
$$\sigma_{15} = (1.4)(1.1)(14.9)^2 = 342$$
$$\sigma_{23} = 231$$
$$\sigma_{24} = 160$$
$$\sigma_{25} = 195$$
$$\sigma_{34} = 260$$
$$\sigma_{35} = 317$$
$$\sigma_{45} = 220$$

These estimates for portfolio inputs are not necessarily the same as would be obtained from historical data. However, the betas for funds 1 and 2 were the historical betas using the prior five years of data. Thus, if the covariance between the residuals for funds 1 and 2 were zero, the estimate of the covariance using the single-index model and the historical estimate would be the same. The covariance between assets 1 and 2 computed directly from the historical data was 271. The estimate from the single-index model was 249. The difference arose because there was a small positive correlation between residuals for fund 1 and fund 2. The justification for using the single-index model to estimate inputs is a belief that this positive residual resulted by chance, and zero is a better estimate of its future value than the actual past value. The optimum proportions using these inputs, a riskless rate of 5%, and the procedures discussed in Chapter 6 are as follows:

FUND A	WITH SHORT SALES	NO SHORT SALES
1	6,926%	78%
2	5,797%	0
3	4,218%	22%
4	−10,143%	0
5	−5,797%	0

The solution with short sales is, of course, unreasonable, both because pension managers cannot short sell and because of the magnitude of the numbers. The large numbers come about because mutual funds are very highly correlated with one another, and small differences result in large positions being taken. In Chapter 9 we analyze the problem when we have developed the tools for a simpler analysis.

QUESTIONS AND PROBLEMS

1. Monthly return data are presented below for each of three stocks and the S&P index (corrected for dividends) for a 12-month period. Calculate the following quantities:

- **A.** alpha for each stock
- **B.** beta for each stock
- **C.** the standard deviation of the residuals from each regression
- **D.** the correlation coefficient between each security and the market
- **E.** the average return on the market
- **F.** the variance of the market

	Security			
Month	A	B	C	S&P
1	12.05	25.20	31.67	12.28
2	15.27	2.86	15.82	5.99
3	−4.12	5.45	10.58	2.41
4	1.57	4.56	−14.43	4.48
5	3.16	3.72	31.98	4.41
6	−2.79	10.79	−0.72	4.43
7	−8.97	5.38	−19.64	−6.77
8	−1.18	−2.97	−10.00	−2.11
9	1.07	1.52	−11.51	3.46
10	12.75	10.75	5.63	6.16
11	7.48	3.79	−4.67	2.47
12	−0.94	1.32	7.94	−1.15

2. A. Compute the mean return and variance of return for each stock in Problem 1 using

(1) The single-index model

(2) The historical data

B. Compute the covariance between each possible pair of stocks using

(1) The single-index model

(2) The historical data

C. Compute the return and standard deviation of a portfolio constructed by placing one-third of your funds in each stock, using

(1) The single-index model

(2) The historical data

D. Explain why the answers to parts A.1 and A.2 were the same, while the answers to parts B.1, B.2, and C.1, C.2 were different.

3. Show that the Vasicek technique leads to a simple proportional weighting of the market beta and the stock's beta if the standard error of all betas is the same.

4. A. If the Blume adjustment equation is fit and the appropriate equation is

$$\beta_{it+1} = 0.41 + 0.60\beta_{i,\,t}$$

what is your best forecast of beta for each of the stocks in Question 1?

B. If the parameters of the Vasicek technique are fit, and they are

$$\beta_1 = 1.00 \qquad \sigma^2_{\bar{\beta}1} = 0.25, \quad \sigma^2_{\bar{\beta}1A} = 0.22, \quad \sigma^2_{\bar{\beta}1B} = 0.36, \quad \sigma^2_{\bar{\beta}1C} = 0.41$$

what is your best forecast of beta for each of the stocks in Question 1?

5.

	Security			
	A	B	C	D
α	2	3	1	4
β	1.5	1.3	0.8	0.9
σ_{ei}	3	1	2	4

Given the preceding data and the fact that $\bar{R}_m = 8$ and $\sigma_m = 5$, calculate the following:

(a) The mean return for each security

(b) The variance of each security's return

(c) The covariance of returns between each security

6. Using the data in Problem 5 and assuming an equally weighted portfolio, calculate the following:

(a) β_p

(b) α_p

(c) σ_p^2

(d) $\bar{R}_p$

7. Using Blume's technique, where $\beta_{i2} = 0.343 + 0.677\beta_{i1}$, calculate β_{i2} for the securities in Problem 5.

8. Suppose $\bar{\beta}_1 = 1$ and $\sigma_{\bar{\beta}1} = 0.25$ $\sigma_{\beta A} = 0.21$ $\sigma_{\beta B} = 0.32$ $\sigma_{\beta C} = 0.18$ $\sigma_{\beta D} = 0.20$, forecast each security's beta using the Vasicek technique.

BIBLIOGRAPHY

1. Alexander, Gordon J., and Benston, P. George. "More on Beta as a Random Coefficient," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1982), pp. 27–36.
2. Alexander, Gordon J., and Chervany, Norman L. "On the Estimation and Stability of Beta," *Journal of Financial and Quantitative Analysis*, **XV,** No. 1 (March 1980), pp. 123–138.
3. Ali, Mukhtar M., and Giaccotto, Carmelo. "Optimum Distribution-Free Tests and Further Evidence of Heteroscedasticity in the Market Model," *Journal of Finance*, **37,** No. 5 (Dec. 1982), pp. 1247–1258.
4. Beaver, W., Kettler, P., and Scholes, M. "The Association between Market Determined and Accounting Determined Risk Measures," *The Accounting Review*, **45** (Oct. 1970), pp. 654–682.
5. Bick, Avi. "On Viable Diffusion Price Processes of the Market Portfolio," *Journal of Finance*, **45,** No. 2 (June 1990), pp. 673–680.
6. Bildersee, John S., and Roberts, Gordon S. "Beta Instability When Interest Rate Levels Change," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 3 (Sept. 1981), pp. 375–380.
7. Blume, Marchall. "Portfolio Theory: A Step toward Its Practical Application," *Journal of Business*, **43,** No. 2 (April 1970), pp.152–173.
8. ——. "On the Assessment of Risk," *Journal of Finance*, **VI,** No. 1 (March 1971), pp. 1–10.
9. ——. "Betas and Their Regression Tendencies," *Journal of Finance*, **X,** No. 3 (June 1975), pp. 785–795.
10. Brenner, Menachem, and Smidt, Seymour. "A Simple Model of Non-stationarity of Systematic Risk," *Journal of Finance*, **XII,** No. 4 (Sept. 1977), pp. 1081–1092.
11. Brown, Stephen. "Heteroscedasticity in the Market Model: A Comment on [61]," *Journal of Business*, **50,** No. 1 (January 1977), pp. 80–83.
12. Chan, Louis, K. C. "The Risk and Return from Factors," *Journal of Financial and Quantitative Analysis*, **33,** No. 2 (June 1998), pp. 159–189.
13. ——. "An Examination of Risk-Return Relationship in Bull and Bear Markets Using Time-Varying Betas," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 2 (June 1982), pp. 265–286.
14. Cohen, K., Maier, S., Schwartz, R., and Whitecomb, D. "The Returns Generation Process, Returns Variance, and the Effect of Thinness in Security Markets," *Journal of Finance*, **XIII,** No. 1 (March 1978), pp. 149–167.
15. Cohen, Kalman, Ness, Walter, Okuda, Hitashi, Schwartz, Robert, and Whitcomb, David. "The Determinants of Common Stock Returns Volatility: An International Comparison," *Journal of Finance*, **XI,** No. 2 (May 1976), pp. 733–740.
16. Cooley, P., Roenfeldt, R., and Modani, N. "Interdependence of Market Risk Measures," *Journal of Business*, **50,** No. 3 (July 1977), pp. 356–363.
17. Cornell, Bradford, and Dietrich, Kimball. "Mean-Absolute-Deviation versus Least-Squares Regression Estimation of Beta Coefficients," *Journal of Financial and Quantitative Analysis*, **XIII,** No. 1 (March 1978), pp. 123–131.
18. Dimson, Elroy, and Marsh, P. "The Stability of UK Risk Measures and the Problem in Thin Trading," *Journal of Finance*, **38,** No. 3 (June 1983), pp. 753–784.
19. Elton, Edwin J., Gruber, Martin J., and Urich, Thomas. "Are Betas Best?" *Journal of Finance*, **XIII,** No. 5 (Dec. 1978), pp. 1375–1384.
20. Fabozzi, Frank, and Francis, Clark. "Stability Tests for Alphas and Betas over Bull and Bear Market Conditions," *Journal of Finance*, **XII,** No. 4 (Sept. 1977), pp. 1093–1099.
21. Fama, Eugene. "Risk, Return, and Equilibrium: Some Clarifying Comments," *Journal of Finance*, **23** (March 1968), pp. 29–40.
22. Fouse, W., Jahnke, W., and Rosenberg, B. "Is Beta Phlogiston?" *Financial Analysts Journal*, **30,** No. 1 (Jan.–Feb. 1974), pp. 70–80.

23. Francis, Jack Clark. "Intertemporal Differences in Systematic Stock Price Movements," *Journal of Financial and Quantitative Analysis*, **X,** No. 2 (June 1975), pp. 205–219.
24. Gonedes, Nicholas J. "Evidence on the Information Content of Accounting Numbers: Accounting-based and Market-based Estimates of Systematic Risk," *Journal of Financial and Quantitative Analysis*, **8** (June 1973), pp. 407–443.
25. Hamada, S. Robert. "The Effect of the Firm's Capital Structure on the Systematic Risk of Common Stocks," *Journal of Finance*, **VII,** No. 2 (May 1971), pp. 435–452.
26. Handa, Puneet, Kothari, S. P., and Wasley, Charles. "The Relation between the Return Interval and Betas: Implications for the Size Effect," *Journal of Financial Economics*, **23,** No. 1 (June 1989), pp. 79–101.
27. Hawawini, Gabriel A. "Intertemporal Cross-Dependence in Securities Daily Returns and the Short-Run Intervaling Effect on Systematic Risk," *Journal of Financial and Quantitative Analysis*, **XV,** No. 1 (March 1980), pp. 139–150.
28. Hawawini, Gabriel A., and Vora, Ashok. "Evidence of Intertemporal Systematic Risks in the Daily Price Movements of NYSE and AMEX Common Stocks," *Journal of Financial and Quantitative Analysis*, **XV,** No. 2 (June 1980), pp. 331–340.
29. Hawawini, Gabriel A., Michel, Pierre A., and Corhay, Albert. "New Evidence on Beta Stationarity and Forecasting for Belgium Common Stocks," *Journal of Business Finance*, **9,** No. 4 (Dec. 1985), pp. 553–560.
30. Hill, Ned C., and Stone, Bernell K. "Accounting Betas, Systematic Operating Risk, and Financial Leverage: A Risk-Composition Approach to the Determinants of Systematic Risk," *Journal of Financial and Quantitative Analysis*, **XV,** No. 3 (Sept. 1980), pp. 595–638.
31. Jacob, Nancy. "The Measurement of Systematic Risk for Securities and Portfolios: Some Empirical Results," *Journal of Financial and Quantitative Analysis*, **VI,** No. 2 (March 1971), pp. 815–833.
32. Joehnk, Michael, and Nielson, James. "The Effects of Conglomerate Merger Activity on Systematic Risk," *Journal of Financial and Quantitative Analysis*, **IX,** No. 2 (March 1974), pp. 215–225.
33. Klemkosky, Robert, and Martin, John. "The Effect of Market Risk on Portfolio Diversification," *Journal of Finance*, **X,** No. 1 (March 1975), pp. 147–153.
34. ——. "The Adjustment of Beta Forecasts," *Journal of Finance*, **X,** No. 4 (Sept. 1975), pp. 1123–1128.
35. Latane, Henry, Tuttle, Don, and Young, Allan. "How to Choose a Market Index," *Financial Analysis Journal*, **27,** No. 4 (Sept.–Oct. 1971), pp. 75–85.
36. Levy, Haim. "Measuring Risk and Performance over Alternative Investment Horizons," *Financial Analysts Journal*, **40,** No. 2 (March/April 1984), pp. 61–67.
37. Levy, Robert. "On the Short-Term Stationary of Beta Coefficients," *Financial Analysts Journal*, **27,** No. 5 (Dec. 1971), pp. 55–62.
38. ——. "Beta Coefficients as Predictors of Return," *Financial Analysts Journal*, **30,** No. 1 (Jan.–Feb. 1974), pp. 61–69.
39. Logue, Dennis, and Merville, Larry. "Financial Policy and Market Expectations," *Financial Management*, **1** (Summer 1972), pp. 37–44.
40. Martin, J., and Klemkosky, R. "Evidence of Heteroscedasticity in the Market Model," *Journal of Business*, **48,** No. 1 (Jan. 1975), pp. 81–86.
41. Officer, R. R. "The Variability of the Market Factor of the New York Stock Exchange," *Journal of Business*, **46,** No. 3 (July 1973), p. 434–453.
42. Pogue, Gerald, and Solnik, Bruno. "The Market Model Applied to European Common Stocks: Some Empirical Results," *Journal of Financial and Quantitative Analysis*, **IX,** No. 6 (Dec. 1974), pp. 917–944.
43. Robichek, Alexander, and Cohn, Richard. "The Economic Determinants of Systematic Risk," *Journal of Finance*, **XXIX,** No. 2 (May 1974), pp. 439–447.
44. Roenfeldt, R., Griepentrog, G., and Pflaum, C. "Further Evidence on the Stationarity of Beta Coefficients," *Journal of Financial and Quantitative Analysis*, **XIII,** No. 1 (March 1978), pp. 117–121.
45. Roll, Richard. "Bias in Fitting the Sharpe Model to Time Series Data," *Journal of Financial and Quantitative Analysis*, **IV,** No. 3 (Sept. 1969), pp. 271–289.

46. Rosenberg, Barr, and Guy, James. "Prediction of Beta from Investment Fundamentals," *Financial Analysts Journal*, **32,** No. 3 (May–June 1976), pp. 60–72.
47. ——. "Prediction of Beta from Investment Fundamentals: Part II," *Financial Analysts Journal*, **32,** No. 3 (July–Aug. 1976), pp. 62–70.
48. Rosenberg, Barr, and Marathe, Vinary. "The Prediction of Investment Risk: Systematic and Residual Risk," Reprint 21, Berkley Working Paper Series (1975).
49. Rosenberg, Barr, and McKibben, Walt. "The Prediction of Systematic and Specific Risk in Common Stocks," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 2 (March 1973), pp. 317–333.
50. Rudd, Andrew, and Rosenberg, Barr. "The 'Market Model' in Investment Management," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 597–606.
51. Schafer, Stephen, Brealey, Richard, and Hodges, Steward. "Alternative Models of Systematic Risk," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1976).
52. Scholes, M., and Williams, J. "Estimating Betas from Non-synchronous Data," *Journal of Financial Economics*, **5,** No. 3 (Dec. 1977), pp. 309–328.
53. Scott, Elton, and Brown, Stewart. "Biased Estimators and Unstable Betas," *Journal of Finance*, **35,** No. 1 (March 1980), pp. 49–56.
54. Sharpe, William. "Mean-Absolute-Deviation Characteristic Lines for Securities and Portfolios," *Management Science*, **18,** No. 2 (Oct. 1971), pp. B1–B13.
55. Sunder, Shyam. "Stationarity of Market Risk: Random Coefficients Tests for Individual Stocks," *Journal of Finance*, **35,** No. 4 (Sept. 1980), pp. 883–896.
56. Theobald, Michael. "Beta Stationarity and Estimation Period: Some Analytical Results," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 5 (Dec. 1981), pp. 747–758.
57. Thompson, Donald, II. "Sources of Systematic Risk in Common Stocks," *Journal of Business*, **40,** No. 2 (April 1978), pp. 173–188.
58. Vasicek, Oldrich. "A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas," *Journal of Finance*, **VIII,** No. 5 (Dec. 1973), pp. 1233–1239.
59. Young, S. David, Berry, Michael A., Harvey, David W., and Page, John R. "Macroeconomic Forces, Systematic Risk, and Financial Variables: An Empirical Investigation," *Journal of Financial and Quantitative Analysis*, **26,** No. 4 (Dec. 1991), pp. 559–565.

8

The Correlation Structure of Security Returns—Multi-Index Models and Grouping Techniques

In Chapter 7 we argued that because of both the huge number of forecasts required and the necessary restrictions on the organizational structure of security analysts, it was not feasible for analysts to directly estimate correlation coefficients. Instead, some structural or behavioral model of how stocks move together should be developed. The parameters of this model can be estimated either from historical data or by attempting to get subjective estimates from security analysts. We have already examined one such model, the single-index model, which assumes that stocks move together only because of a common comovement with the market. Two other approaches have been widely used to explain and estimate the correlation structure of security returns: multi-index models and averaging techniques.

Multi-index models are an attempt to capture some of the nonmarket influences that cause securities to move together. The search for nonmarket influences is a search for a set of economic factors or structural groups (industries) that account for common movement in stock prices beyond that accounted for by the market index itself. Although it is easy to find a set of indexes that is associated with nonmarket effects over any period of time, as we will see, it is quite another matter to find a set that is successful in predicting covariances that are not market related.

Averaging techniques are at the opposite end of the spectrum from multi-index models. Multi-index models introduce extra indexes in the hope of capturing additional information. The cost of introducing additional indexes is the chance that they are picking up random noise rather than real influences. Averaging techniques smooth the entries in the historical correlation matrix in an attempt to "damp out" random noise and so produce better forecasts. The potential disadvantage of averaging models is that real information may be lost in the averaging process.

In this chapter we examine both multi-index models and averaging models. Several of the models put forth in the finance literature are discussed, as are some of the empirical evidence on their relative merits.

At this point, we should mention that there are other uses for multi-index models besides predicting correlation coefficients. Multi-index models can be used to form expectations about returns and study the impact of events, as a method for tailoring the return distribution of a portfolio to the specific needs of an investor, and as a method for attributing the

cause of good or bad performance on a portfolio. These are subjects to which we will return later in the book. However, the reader should be alerted to these other possible uses. We will close this chapter with a discussion of some multi-index models using fundamental data that have recently been developed as a step toward building a general equilibrium model of security returns. We return to this class of model in Chapter 16.

MULTI-INDEX MODELS

The assumption underlying the single-index model is that stock prices move together only because of common movement with the market. Many researchers have found that there are influences beyond the market that cause stocks to move together. For example, as early as 1966, King (1966) presented evidence on the existence of industry influences. Two different types of schemes have been put forth for handling additional influences. We have called them the general multi-index model and the industry index model.

General Multi-index Models

Any additional sources of covariance among securities can be introduced into the equations for risk and return simply by adding these additional influences to the general return equation. Let us hypothesize that the return on any stock is a function of the return on the market, changes in the level of interest rates, and a set of industry indexes. If R_i is the return on stock i, then the return on stock i can be related to the influences that affect its return in the following way:

$$R_i = a_i^* + b_{i1}^* I_1^* + b_{i2}^* I_2^* + \cdots + b_{iL}^* I_L^* + c_i$$

In this equation I_j^* is the actual level of index j and b_{ij}^* is a measure of the responsiveness of the return on stock i to changes in the index j. Thus b_{ij}^* has the same meaning as β_i in the case of the single-index model. A b_{ij}^* of 2 would mean that if the index increased (decreased) by 1%, the stock's return is expected to increase (decrease) by 2%. As in the case of the single-index model, the return of the security not related to indexes is split into two parts: a_i^* and c_i where a_i^* is the expected value of the unique return. This is the same meaning it had in the single-index model. Variable c_i is the random component of the unique return; it has a mean of zero and a variance we will designate as σ_{ci}^2.

Although a multi-index model of this type can be employed directly, the model would have some very convenient mathematical properties if the indexes were uncorrelated (orthogonal). This would allow us to simplify both the computation of risk and the selection of optimal portfolios. Fortunately, this presents no theoretical problems because it is always possible to take any set of correlated indexes and convert them into a set of uncorrelated indexes. The method for doing so is outlined in Appendix A. Using this methodology, the equation can be rewritten as[1]

$$R_i = a_i + b_{i1} I_1 + b_{i2} I_2 + b_{i3} I_3 + \cdots + b_{iL} I_L + c_i$$

where all I_j are uncorrelated with each other. The new indexes still have an economic interpretation. Assume I_1^* was a stock market index and I_2^* an index of interest rates. I_2 is now an index of the difference between actual interest rates and the level of interest rates that

[1]The asterisks have been removed to indicate that the indexes and coefficients are now different. Actually, if the procedure in Appendix A at the end of this chapter is followed, $I_1 = I_1^*$, but all others are different. In applications it may be easier for analysts to estimate the model with correlated indexes. This model can then be transformed into one with uncorrelated indexes for purposes of portfolio selection.

would be expected given the rate of return on the stock market (I_1). Similarly, b_{i2} becomes a measure of the sensitivity of the return on stock i to this difference. We can think of b_{i2} as the sensitivity of stock i's return to a change in interest rates when the rate of return on the market is fixed.

Not only is it convenient to make the indexes uncorrelated, but it is also convenient to have the residual uncorrelated with each index. Formally, this implies that $E[c_i(I_j - \bar{I}_j)] = 0$ for all j. The implication of this construction is that the ability of Equation (8.1) to describe the return on any security is independent of the value any index happens to assume. When the parameters of this model are estimated via regression analysis, as is usually done, this will hold over the period of time to which the model is fitted.

The standard form of the multi-index model can be written as follows:

BASIC EQUATION:

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + b_{i3}I_3 + \cdots + b_{iL}I_L + c_i$$

$$\text{for all stocks } i = 1, ..., N \qquad (8.1)$$

BY DEFINITION

1. Residual variance of stock i equals σ_{ci}^2, where $i = 1, ..., N$.

2. Variance of index j equals σ_{Ij}^2, where $j = 1, ..., L$.

BY CONSTRUCTION

1. Mean of c_i equals $E(c_i) = 0$ for all stocks, where $i = 1, ..., N$.

2. Covariance between indexes j and k equals $E[(I_j - \bar{I}_j)(I_k - \bar{I}_k)] = 0$ for all indexes, where $j = 1, ..., L$ and $k = 1, ..., L$ $(j \neq k)$.

3. Covariance between the residual for stock i and index j equals $E[c_i(I_j - \bar{I}_j)] = 0$ for all stocks and indexes, where $i = 1, ..., N$ and $j = 1, ..., L$.

BY ASSUMPTION

1. Covariance between c_i and c_j is zero $(E(c_ic_j) = 0)$ for all stocks where $i = 1, ..., N$ and $j = 1, ..., N$ $(j \neq i)$.

The assumption of the multi-index model is that $E(c_ic_j) = 0$. This assumption implies that the only reason stocks vary together is because of common comovement with the set of indexes that have been specified in the model. There are no factors beyond these indexes that account for comovement between any two securities. There is nothing in the estimation of the model that forces this to be true. This is a simplification that represents an approximation to reality. The performance of the model will be determined by how good this approximation is. This, in turn, will be determined by how well the indexes that we have chosen to represent comovement really capture the pattern of comovement among securities.

The expected return, variance, and covariance among securities when the multi-index model describes the return structure are derived in Appendix B and are equal to the following:

1. Expected return is

$$\bar{R}_i = a_i + b_{i1}\bar{I}_1 + b_{i2}\bar{I}_2 + \cdots + b_{iL}\bar{I}_L \qquad (8.2)$$

2. Variance of return is

$$\sigma_i^2 = b_{i1}^2\sigma_{I1}^2 + b_{i2}^2\sigma_{I2}^2 + \cdots + b_{iL}^2\sigma_{IL}^2 + \sigma_{ci}^2 \qquad (8.3)$$

3. Covariance between security i and j is

$$\sigma_{ij} = b_{i1}b_{j1}\sigma^2_{I1} + b_{i2}b_{j2}\sigma^2_{I2} + \cdots + b_{iL}b_{jL}\sigma^2_{IL} \tag{8.4}$$

From Equations (8.2), (8.3), and (8.4) it is clear that the expected return and risk can be estimated for any portfolio if we have estimates of a_i for each stock, and estimates of b_{ik} for each stock with each index, an estimate of σ^2_{ci} for each stock and, finally, an estimate of the mean ($\bar{I}_j$) and variance σ^2_{Ij} of each index. This is a total of $2N + 2L + LN$ estimates. For an institution following between 150 and 250 stocks and employing 10 indexes, this calls for between 1,820 and 3,020 inputs. This is larger than the number of inputs required for the single-index model but considerably less than the inputs needed when no simplifying structure was assumed. Notice that now analysts must be able to estimate the responsiveness of each stock they follow to several economic and industry influences.

This model can also be used if analysts supply estimates of the expected return for each stock, the variance of each stock's returns, each index loading (b_{ik} between each stock i and each index k), and the means and variances of each index. This is the same number of inputs ($2N + 2L + LN$). However, the inputs are in more familiar terms. As discussed at several points in this book, the inputs needed to perform portfolio analysis are expected returns, variances, and correlation coefficients. By having the analysts estimate means and variances directly, it is clear that the only input derived from the estimates of the multi-index models is correlation coefficients. We stress this point because later in this chapter, we evaluate the ability of a multi-index model to aid in the selection of securities by examining its ability to forecast correlation coefficients.

There is a certain type of multi-index model that has received a large amount of attention. This class of models restricts attention to market and industry influences. Alternative industry index models result from different assumptions about the behavior of returns and, hence, differ in the type and amount of input data needed. We now examine these models.

Industry Index Models

Several authors have dealt with multi-index models that start with the basic single-index model and add indexes to capture industry effects. The early precedent for this work can be found in King (1966), who measured effects of common movement between securities beyond market effects and found that this extra market covariance was associated with industries. For example, two steel stocks had positive correlation between their returns, even after the effects of the market had been removed.[2]

If we hypothesize that the correlation between securities is caused by a market effect and industry effects, our general multi-index model could be written as

$$R_i = a_i + b_{im}I_m + b_{i1}I_1 + b_{i2}I_2 + \cdots + b_{iL}I_L + c_i$$

where

- I_m is the market index
- I_j are industry indexes that are constrained to be uncorrelated with the market and uncorrelated with each other

[2]King (1966) found that over the entire period studied, 1927–1960, about half of the total variation in a stock's price was accounted for by a market index, while an average of another 10% was accounted for by industry factors. In the latter part of the period he studied, the importance of the market factor dropped to 30%, while the industry factors continued to explain 10% of price movement.

The assumption behind this model is that a firm's return can be affected by the market plus several industries. For some companies this seems appropriate as their lines of business span several traditional industries. However, some companies gain the bulk of their return from activities in one industry and, perhaps of more importance, are viewed by investors as members of a particular industry. In this case, the effects on the firm's return of indexes for industries to which they do not belong are likely to be small, and their inclusion may introduce more random noise into the process than the information they supply. This has prompted some authors to advocate a simpler form of the multi-index model: one that assumes that returns of each firm are affected only by a market index and one industry index. Furthermore, the model assumes that each industry index has been constructed to be uncorrelated with the market and with all other industry indexes. For firm i in industry j, the return equation can be written as

$$R_i = a_i + b_{im} I_m + b_{ij} I_j + c_i$$

The covariance between securities i and k can be written as

$$b_{im} b_{km} \sigma_m^2 + b_{ij} b_{kj} \sigma_{Ij}^2$$

for firms in the same industry and as

$$b_{im} b_{km} \sigma_m^2$$

for firms in different industries. Notice that the number of inputs needed for portfolio selection has been cut to $4N + 2L + 2$.

The data needed are the expected return and variance for each stock, the loading of each stock on the market and industry index, and, finally, the mean and variance of each industry index and the market index.[3]

How Well Do Multi-index Models Work?

At this point it is worth examining how well these multi-index models have performed when the parameters are estimated from historical data.[4] Remember, multi-index models lie in an intermediate position between the full historical correlation matrix itself and the single-index model in ability to reproduce the historical correlation matrix. The more indexes added, the more complex things become and the more accurately the historical correlation matrix is reproduced. However, this does not imply that future correlation matrices will be forecast more accurately. Because there are an infinite number of multi-index models that can be tried, one cannot unequivocally say that multi-index models are better or worse than single-index models. However, we can examine some typical results on several multi-index models to see how well they work.

---

[3]As the reader can imagine, there is more than one way to write any model. This particular type of multi-index model has been popularized in another form by Cohen and Pogue (1967).

It can be shown that the model Cohen and Pogue call the diagonal form of the multi-index model is identical to the form we have been discussing. The advantage of expressing the input data as suggested by Cohen and Pogue is that the analyst can deal directly with responsiveness of industries to the market. This may be easier than dealing with the responsiveness of stocks to industry indexes with market influences removed.

[4]All of the tests of the various models we've discussed have estimated the models using historical data. There is no reason that the estimates could not come from analysts. The ability of analysts to make these estimates and their value is still an open question.

Let us start with the most general form of the multi-index model:

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + b_{i3}I_3 + \cdots + b_{iL}I_L + c_i$$

This model explains firm returns in terms of a set of uncorrelated indexes.[5]

Before discussing the results, it is worth digressing for a moment to see how we might judge the performance of these models. Remember that all index models lead to the same estimates of expected returns and a stock's own variance (as opposed to covariances) when estimated from historical returns and variances. Furthermore, if analysts are used to estimating expected return and variance, the only estimate from a model is an estimate of the covariance. However, the covariance is the product of standard deviations and correlation coefficients. If analysts are used to estimating standard deviations, any differences in performance that exist must arise from differences in estimating the correlation structure of security returns. The most direct test of alternative models is to examine how well they estimate the future correlation matrix of security returns. Differences between forecasts and actual results can be measured, and the statistical significance of these differences can be judged. While tests of statistical significance are useful for judging the superiority of forecasting techniques, tests of economic significance are often of more interest. Tests of economic significance examine the difference in return or profit that results from basing forecasts on one technique rather than on another. In this case, the future returns (at alternative specified risk levels) that would result from selecting portfolios based on each forecasting model can be examined.

The results of any testing of how well a multi-index model performs in forecasting the future depends on how the indexes are defined. The simplest approach, and one that is widely used in finance, is to let the data define the indexes. There is a standard statistical technique, called principal components analysis, that extracts from past values of the variance–covariance matrix a set of indexes that best explain (reproduce) the historical matrix itself. Elton and Gruber (1973) performed extensive tests on indexes derived from the historical correlation matrix. They found (that on both statistical grounds and economic grounds) adding additional indexes derived from the past correlation matrix to the single–index model led to a decrease in performance. Although adding more indexes led to a better explanation of the historical correlation matrix, it led both to a poorer prediction of the future correlation matrix and to the selection of portfolios that, at each risk

[5]A mathematical technique exists that allows a set of indexes that meets the criteria for this model to be constructed from a set of returns. The technique is called *principal components analysis*. Principal components analysis will extract from a historical variance–covariance matrix of returns that index (weighting of the individual returns) that best explains (reproduces) the variance of the original data. This index is called the first principal component. Principal components analysis then proceeds to extract the index that explains as much as possible of the variance of the original data unexplained by the first principal component, given that this second index is constrained to be uncorrelated with the first index. It proceeds to sequentially form additional indexes, ensuring that each index formed explains as much as possible of the variation in the data that has not been explained by previous indexes, given that each index extracted is uncorrelated with each index previously extracted. This technique can be used until the number of indexes extracted equals the number of stocks whose variance–covariance matrix is being examined. At this point, the principal components can exactly reproduce the historical variance–covariance matrix. However, because the first principal component explains as much as possible of the historical variance–covariance matrix, the second explains as much as possible of the remaining variance, and so on, we would expect the last few principal components to have almost no explanatory power. In fact, to the extent that there is any real underlying structure to the data, most of the correlation matrix should be explained by the first few principal components.

Elton and Gruber (1973) used principal components analysis on 76 firms and found that the percentage of the variance in the original data explained with 1, 3, 8, and 17 principal components was 36%, 45%, 61%, and 75%, respectively.

level, tended to have lower returns. In short, these added indexes introduced more random noise than real information into the forecasting process. We review some evidence later in this chapter.

The evidence that a generalized multi-index model, where the indexes are extracted according to explanatory power from past data, does not perform as well as a single-index model is very strong. This does not imply that a different form of a multi-index model might not work better than a single-index model. Indexes based on interest rates or oil prices or other fundamental factors affecting different companies in different ways may lead to better performance. One would expect that other influences exist that should have a major and lasting influence on the correlation structure of stock prices. Whether they do is a matter for empirical research.

Another test of the multi-index model was performed by Cohen and Pogue (1967). They examined the use of a specialized multi-index model to select portfolios (test of economic significance).[6] Standard industrial classifications were used to divide the stocks in their sample into industries. Standard industrial classifications group firms by end product such as steel or chemical. Single-index models and a multi-index model, with a market and industry index, were then run. While Cohen and Pogue tested results, both over the period to which the models were fit and over the forecast period, only the latter set of tests is of interest to a person considering the adoption of these models. Cohen and Pogue conclude that with respect to these tests, the single-index model has more desirable properties. The single-index model led to lower expected risks and is much simpler to use.[7]

Whereas Cohen and Pogue accepted standard industrial classifications in their analysis, other authors have sought to employ industry index models where industries were defined not in terms of a standard classification but in terms of the tendency of firms to act alike. Procedures for forming homogeneous groups of firms or pseudo-industries were first examined in Elton and Gruber (1970) and later again in Elton and Gruber (1973) and Elton, Gruber, and Blake (1999). Pseudo-industries are formed simply by combining firms whose returns are highly correlated into an industry.[8] Once pseudo-industries are formed, an index can be calculated to represent the return on each pseudo-industry. The good news in developing pseudo-industries is that they seem to be fairly stable over time. However, despite this, and despite their intuitive appeal, their performance in a multi-index model does not seem to be better than traditional industry indexes.

There has been a renewed interest in multi-index models. The testing has been to see how many indexes best explain the historical variance–covariance or correlation matrix. Roll and Ross (1980) report that at least three indexes are needed to explain the historical variance–covariance matrix. Dhrymes, Friend, and Gultekin (1984) show that the number of indexes that are needed is very dependent on the number of firms that are being analyzed. Depending on the sample size, they find that many more than three are needed. Finally, Gibbons (1982), analyzing bond and stock data, finds that six or seven indexes are needed. Chen, Roll, and Ross (1986), and Burmeister writing with others (1986, 1987, 1988), have produced a set of multi-index models based on a priori hypothesized set of macroeconomic variables. Fama and French (1993) have proposed a set of indexes based

[6]The specialized form of the model they tested was their diagonal form of the multi-index model.

[7]Cohen and Pogue (1967) also tested a more elaborate form of the multi-index industry model. In this form the entire covariance structure between industry indexes was employed. However, the performance of this model was inferior to the simpler diagonal form of the industry multi-index model and hence inferior to the single-index model.

[8]The correlation has been examined both for raw returns and for returns after the market return has been removed.

on firm characteristics. Because of the growing importance of these latter two types of models, we devote a special section to their description at the end of this chapter. Those models are extremely interesting and have several potential applications in finance.

Up to this point, we have discussed the use of multi-index models based on the historic correlation data as a way of forecasting the future correlation structure between security returns. While this use holds great promise for the future, the results, to date, have been mixed. A natural question arises: if the addition of more indexes to a single-index model can, at times, introduce more random noise than real information into the forecasting process, might not a technique that smooths more of the historical data lead to better results?

AVERAGE CORRELATION MODELS

The idea of averaging (smoothing) some of the data in the historical correlation matrix as a forecast of the future has been tested by Elton and Gruber (1973) and Elton, Gruber, and Urich (1978).

The most aggregate type of averaging that can be done is to use the average of all pairwise correlation coefficients over some past period as a forecast of each pairwise correlation coefficient for the future. This is equivalent to the assumption that the past correlation matrix contains information about what the average correlation will be in the future but no information about individual differences from this average. This model can be thought of as a naive model against which more elaborate models should be judged. We refer to this model as the overall mean model.

A more disaggregate averaging model would be to assume that there was a common mean correlation within and among groups of stocks. For example, if we were to employ the idea of traditional industries as a method of grouping, we would assume that the correlation between any two steel stocks was the same as the correlation between any other two steel stocks and was equal to the average historical correlation among steel stocks. The averaging is done across all pairwise correlations among steel stocks in a historical period. Similarly, the correlation among any steel stocks and any chemical stocks is assumed to be equal to the correlation between any other steel stock and any other chemical stock and is set equal to the average of the correlations between each chemical and each steel stock. When this is done, with respect to traditional industry classifications, it will be referred to as the traditional mean model. The same technique has been used by Elton and Gruber (1973) with respect to pseudo-industries.

The overall mean has been extensively tested against single-index models, general multi-index models, and the historical correlation matrix itself. Tests have been performed using three different samples of stocks over a total of four different time periods. In every case, the use of the overall mean model outperformed the single-index model, the multi-index model, and the historical correlation matrix. The differences in forecasting future correlation coefficients were almost always statistically significant at the 0.05 level. Furthermore, for most risk levels, the differences in portfolio performance were large enough to have real economic significance. Using the overall mean technique, as opposed to the best of the single-index model, the multi-index model, or the historical correlation model, often led to a 25% increase in return (holding risk constant).

The next logical question is what happens when we introduce some disaggregation into the results by using the traditional mean or pseudo-mean model. Here the results are much more ambiguous. Averaging models based on either traditional industries or pseudo-industries outperformed single-index models, multiple-index models, and the historical correlation matrix both on statistical and economic criteria. However, their differences

from each other and from the overall mean were much less clear. The ordering of these three techniques was different over different time periods and at different risk levels in the same time period. At this point all we can say is that, although it is worth continuing to investigate the performance of traditional mean and pseudo-mean averaging models, their superiority over the overall mean model has not yet been demonstrated.

MIXED MODELS

Another model that has received attention is a combination of the models discussed in Chapter 7 and those introduced in this chapter. We call them *mixed models*. In a mixed model, the single-index model is used as the basic starting point. However, rather than assume that the extramarket covariance is zero, a second model is constructed to explain extramarket covariance. This concept should not be new to the reader. If we consider a general multi-index model where the first index is the market, then we can consider all other indexes as indexes of extramarket covariance. What is new is the way that extramarket covariance is predicted. The most widely known model of this type is that described by Rosenberg (1974). In Chapter 7 we discussed Rosenberg's methods of relating beta to a set of fundamental and technical data. Rosenberg has used the same method for predicting extramarket covariance. He relates extramarket covariance to the same type of fundamental variables and industry membership coefficients that were discussed in Chapter 7. After removing the market index, he regresses the extramarket covariance on 114 variables. These variables include traditional industry classification as well as firm variables such as debt–equity ratios and dividend payout measures. Initial results with this type of analysis appear quite promising, although extensive tests of the forecast ability have not been performed.

Another approach worth exploring is to apply the same type of averaging techniques discussed earlier directly to the extramarket covariance. That is, instead of performing the averaging on the correlation coefficients themselves, perform the averaging on the correlations of the residuals from the single-index model. For example, a traditional industry averaging scheme might be used. In this case, after removing the market influence, the residuals for each stock could be averaged within and between industries. Then the correlation between any two stocks would be predicted by combining their predicted correlation from the single-index model with the extramarket correlation predicted from the averaging model.

FUNDAMENTAL MULTI-INDEX MODELS

Two types of fundamental multi-index models have received a great deal of attention in the academic and practitioner literature. One set of models stems from the work of Fama and French (1993). The other stems from the work of Chen, Roll, and Ross (1986).

Fama–French Models

Fama and French laid the basis for a multi-index model based on firm characteristics in a series of articles published in the early 1990s. They found that both size (market capitalization) and the ratio of book value of equity to the market value of equity have a strong role in determining the cross section of average return on common stocks. Reasoning that both are proxies for risk, they found (in multivariate tests) that a cross section of average returns is negatively related to size and positively related to book to market ratios. In simple terms, small firms and firms with low book to market are riskier than other firms.

How, then, do they incorporate these variables into a multi-index time series model of returns? Components of the series, such as the book value of equity, are reported at most four times a year. For time series tests, we need at least monthly observations. Fama and French formulated three indexes to explain the difference between the return on any stock and the riskless rate of interest (30-day Treasury bill rate).

The concept behind the size and book to market indexes is to form portfolios that will have returns that mimic the impact of the variables. By forming portfolios that have observable monthly returns, Fama and French convert a set of variables that cannot be observed at frequent intervals into a set of traded assets that have prices and returns that can be observed at any moment of time and over any interval.

Constructing each of these variables is a two-step process.

Step 1: Size for any firm is defined once a year as the total market value of equity (price times number of shares) as of June. Two groups are defined: one containing all stocks on the New York Stock Exchange (NYSE), AMEX, and NASDAQ that have a size larger than the median size of a stock on the NYSE and a second containing all smaller stocks. The cutoff is chosen from the NYSE rather than from all exchanges to have a reasonable total market capitalization in the smaller half. For example, the lower half of the size category only separated 8% of the market value of all stocks contained in both groups. Unlike the two groups for size, firms were broken into three groups on the basis of the book value of equity to the market value of equity (BE/ME). The break points are defined by the break points of the lowest 30% (S), middle 40% (M), and highest 30% (H) of stocks in the NYSE.

This two-way classification is then used to form five marketable portfolios each year, with the first containing all stocks that fall in the small size low book to market category and the fifth containing the biggest market value and high book to market categories. Returns for the market-weighted portfolios in each of these five categories are estimated.

Step 2: Define the actual indexes used to explain return. The size variable is formulated as small minus big (SMB) and is defined as the difference between two portfolios. The first is the simple average of the returns on the three small portfolios (for the three groups of BE/ME) and the second is the return on the three large portfolios.

The second variable is defined as high minus low (HML) and, using a procedure analogous to that given earlier, represents a series of monthly returns as the high BE/ME portfolio minus the low BE/ME portfolio.

By breaking the portfolios into five groups and then forming two portfolios, the attempt was made to have the size variable free of the book to market effects and the book to market variable free of size effects. That these variables do so can be taken by the fact that the correlation between the size and book to market variables was only -0.08.

Finally, the third variable used is simply the return on the market minus the Treasury bill rate. Note that all of the variables are formulated as zero net investment portfolios. This has implications for equilibrium tests of the model, to which we will return in Chapter 16.

Fama and French show that adding the size and book to equity indexes to the excess return on the market increases the explanatory power of the model. For example, for a portfolio of large size, high BE/ME stocks, the R^2 goes from 0.69 with just the market as the explanatory variable to 0.83 when all three variables are included.[9]

[9]While the Fama–French model is widely used, authors frequently add a fourth variable. Carhart (1997), drawing on the anomaly literature discussed in Chapter 17 on Efficient Markets, finds that momentum is positively related to future returns. Momentum is measured by the return on any stock relative to the return on stocks in general over the past twelve months. More specifically, Carhart formulates the momentum variable as the return on an equally weighted portfolio of the 30% of stocks with the highest past return, minus the return on an equally weighted portfolio of stocks with the lowest return.

Chen, Karceski, and Lakonishok (1999) test the ability of the Fama–French model to produce future correlations against the model that assumes all correlations are the same, the market model, a four-factor extension of the Fama–French model, and index models of larger dimensionality. They find that the constant correlation model produces the lowest forecast error of all the models. They point out that the advantage of the Fama–French model is that it allows the user to explicitly see the effect of the size and BE/ME on correlations.

Chen, Roll, and Ross Model

The second group of fundamental multi-index models of stock returns was published by Chen, Roll, and Ross (1986). Although the purpose of their article was to explain equilibrium returns (a subject we discuss at great length in Chapter 16), their analysis laid the groundwork for many of the models that were to follow. Chen, Roll, and Ross hypothesized a broad set of influences that could affect security returns. Their work is based on two concepts. The first is that the value of a share of stock is equal to the present value of future cash flows to the equity holder. Thus an influence that affects either the size of future cash flows or the function (discount rates) used to value cash flows impacts price. Once a set of variables that affects prices is identified, their second concept comes into play. They argue that because current beliefs about these variables are incorporated in price, only innovations or unexpected changes in these variables can affect return.

In a series of articles, Burmeister, McElroy, and others (1986, 1987, 1988) have continued the development of a multi-index model building on the work of Chen, Roll, and Ross. They find that five variables are sufficient to describe security returns. They employ two variables that are related to the discount rate used to find the present value of cash flows, one related to both the size of the cash flows and discount rates, one related only to cash flows, and a remaining variable that captures the impact of the market not incorporated into the first four variables. Let us briefly discuss each of the variables.

Prices are affected by the rate at which future cash flows are discounted by an investor. They argue that the average rate used depends on two influences. The first depends on how much more an investor requires to buy a more risky instrument rather than a safe one. The second is the shape of the discount function (the rate at which the investor discounts cash flows far in the future versus the rate used to discount near cash flows). Remember, it is unexpected changes or innovations in these variables rather than their levels that affect returns.

The first variable employed by Burmeister et al. is the unexpected difference in return between 20-year government bonds and 20-year corporate bonds. The interest payments on government bonds are considered to be riskless, whereas corporations may default on their payments. Thus return differences in these series measure default risk. They argue that differences in this series from its average value are unexpected. Because the average monthly difference between corporate bonds and government bonds over a long time period is one-half of 1% per month, their first variable is[10]

$$I_1 = \text{one-half of 1\% plus the return on long-term government bonds minus the return on long-term corporate bonds}$$

The second variable measures the shape of the interest rate relationship with maturity. Called *term structure*, it is measured as

[10]The authors use the data of Ibbotson and Sinquefield (1982) for their return series.

I_2 = return on long-term government bonds minus return on the one-month Treasury bill one month in the future

The authors find that this variable has a zero mean and zero autocorrelation and thus argue that any nonzero value is unexpected.[11]

The third variable is a measure of unexpected deflation. To the extent that investors are concerned with real cash flows (cash flows after adjusting for inflation) or adjust discount rates to real values, the rate of deflation should affect stock prices. Thus unexpected changes in deflation should affect returns:

I_3 = rate of inflation expected at the beginning of the month minus the actual rate of inflation realized at the end of the month

The fourth variable uses the unexpected change in the growth rate in real final sales as a proxy for the unexpected changes in long-run profits for the economy:

I_4 = expected long-run growth rate in real final sales expected at the beginning of the month minus the expected long-run growth rate in real final rates expected at the end of the month[12]

To the extent that these four influences do not capture all of the macroeconomic (and psychological) factors affecting stock returns, there may be an impact of the market itself. More specifically, Burmeister et al. wish to examine the impact of the market on stock returns after the influence of their first four variables is removed. To do this, they form a fifth variable. As a proxy for the market, they use the return on the S&P index. The fifth variable is the return on the S&P 500 index, which is uncorrelated with any of the four indexes already discussed. To obtain this variable, they first run a time series regression of the S&P index on the four variables discussed previously and obtain the following results:

$$R_M - R_F = 0.0022 - 1.33I_1 + 0.56I_2 + 2.29I_3 - 0.93I_4$$
$$R_2 = 0.24$$

The author's last variable, I_5, is simply the difference between the excess return on the market for any month and the excess return predicted from the estimated equation or the time series of

$$I_5 = \left(R_m - R_F\right) - \left(0.0022 - 1.33I_1 + 0.56I_2 + 2.29I_3 - 0.93I_4\right)$$

How can we judge whether this model makes sense? If the model is a reasonable return-generating process, we would expect the first four variables to be related to the market in a sensible manner, and we would expect returns on individual stocks to be related to the five variables in a sensible manner. Let us first look at the relationship between the S&P index and the first four variables.

As shown previously, the first four influences account for about 25% of the movement in the S&P index. In addition, the coefficient on each variable is statistically significant at the 5% level and has the sign that theory would lead us to expect.

Consider the second variable, I_2. If the premium for holding longer maturity instruments is high, the rate of return required by the market should also be high, and stock returns should be high. Hence the sign of the coefficient of I_2 should be positive. Similarly, if I_1 is

[11]Data for the first two variables are taken from Ibbotson and Sinquefield (1982).

[12]The third and fourth variables are constructed from the National Income Accounts. Expected inflation is found by time series treatment (Kalman filter) of past inflation series. Expectations are forecast by using a lagged autoregressive model involving lagged values of growth in final sales and growth in disposable income.

Table 8.1 Sector Sensitivities

	I_1 Default	I_2 Term Structure	I_3 Deflation	I_4 Growth	I_5 Residual Market	R^2
Sector name						
Cyclical	−1.63	0.55	2.84	−1.04	1.14	0.77
Growth	−2.08	0.58	3.16	−0.92	1.28	0.84
Stable	−1.40	0.68	2.31	−0.22[a]	0.74	0.73
Oil	−0.63[a]	0.31	2.19[a]	−0.83[a]	1.14	0.50
Utility	−1.06	0.72	1.54	0.23[a]	0.62	0.67
Transportation	−2.07	0.58	4.45	−1.13	1.37	0.66
Financial	−2.48	1.00	3.20	−0.56[a]	0.99	0.72

[a]Indicates *not* statistically different from zero at the 5% level.

large, it indicates a small risk premium is demanded by the market and stock returns should be low. Thus the coefficient I_1 should have a negative sign. I_3 measures deflation. Deflation, I_3, should be and is associated with an increase in stock returns. Thus its sign should be and is positive. The fourth variable measures the decrease in expectations of sales growth. If expectations decrease during a period, prices should drop and returns should be high. Hence the negative relationship found by the authors.[13]

How well does this five-index model explain returns? Fitting the model to 70 firms, the authors find that 215 out of the 350 regression parameters (b_{ij}s) are significantly different from zero at the 5% level (have t values of 1.98 or greater), and the model typically accounts for between 30% and 50% in the variation of the return of individual stocks. Furthermore, they form portfolios of securities in industries or sectors and regress returns on these portfolios against the indexes. The results have intuitive appeal. For example, Berry, Burmeister, and McElroy (1988) examine the sensitivities (b_{ij}s) of seven economic sectors to their five risk indexes. The seven sectors they examine are cyclical, growth, stable, oil, utility, transportation, and financial. These results are shown in Table 8.1. Note that the financial sector has the highest sensitivity (of any of the seven sectors) on I_1 (default risk) and I_2 (term structure). Firms in this industry are highly leveraged, and we would expect their performance to be very sensitive to changes in the term structure or risk structure of interest rates. As another example, utilities have the lowest sensitivity to deflation I_3 and growth in profits I_4. Utilities are governed by rate-of-return regulation and so can pass on much of the impact of deflation and profit changes to their customers in the form of higher (or lower) prices. As a final example, note that the highest sensitivity to the market influence is associated with growth stocks and the lowest sensitivity is utilities. The impact of other influences not captured by the first four indexes is captured in I_5, including market psychology. It seems reasonable that growth stocks are most sensitive to this influence, and utility stocks, which are often described as pseudo bonds, are least sensitive to it.

The model we have just described represents an example of the type of fundamental risk model that is beginning to have an impact on industry as well as the academic profession. A return-generating process developed by Salomon Brothers (1989) is in the spirit of the type of model we have been discussing. This model uses seven variables to explain the return on securities:

[13]Recall that the five variables are supposed to be surprises or innovations, and as such they should not be able to be predicted from their own past values. Burmeister, McElroy, and others test this by examining the time series of the indexes themselves and conclude that they cannot predict the value of the index from their past values (all autocorrelations are close to zero).

1. ***Economic growth.*** As a proxy for long-run growth trends in the economy, it uses year-to-year changes in total industrial production. This series provides a gauge of general economic well-being.[14]
2. ***Business cycle.*** They argue that the shorter-term cyclical behavior of the economy is captured by the difference in return on investment-grade corporate bonds and U.S. Treasuries. They use bonds with about a 20-year maturity. They argue that changes in the spread between the two instruments capture the risk of default.
3. ***Long-term interest rates.*** They argue that changes in the long rate reflect an alteration in the relative attractiveness of financial assets and should induce a change in the portfolio mix. This model uses the yield change in 10-year Treasuries as an indicator of the attractiveness of default-free bonds.
4. ***Short-term interest rates.*** Similarly, a change in short-term interest rates would alter the supply of assets for investment in longer-term instruments, such as stocks and bonds. The model uses the yield change in one-month U.S. Treasury bills as an indicator of changes at the short end of the yield curve.
5. ***Inflation stock.*** The Consumer Price Index (CPI) is used to measure inflation. The stock element is measured as the difference between realized inflation and expected inflation.[15]
6. ***U.S. dollar.*** The impact of currency fluctuations on the stock market is measured by changes in a 15-country, trade-weighted basket of currencies. Salomon finds a statistically stable relationship between returns on stocks and currency fluctuations.
7. That part of the market index that is uncorrelated with the six indexes previously described.

Salomon Brothers has been employing their multi-index model for some time. They report that using monthly data, this model explains on average 41% of the fluctuations in return for individual stocks contained in a sample of 1,000 institutional-quality stocks. Models of this type are most promising. We will return to examine this again when we discuss equilibrium prices in Chapter 16.

Improving Forecasts of Correlation

There has been a recent renewed interest in predicting correlation and covariance. More and more, industry and the academic profession have come to realize how difficult it is to estimate, based on judgment, the correlation among stocks. Recent literature has compared and tested several of the techniques we have already discussed, derived new groupings of firms and forecasts assuming the correlation constant within and among groups, and examined methods for combining estimates from different techniques into better estimates.

For an excellent discussion comparing many of the alternative forecasting techniques we have presented earlier in this chapter, see Chan, Karceski, and Lakonishok (1999) and Elton, Gruber, and Spitzer (2006). Consistent with early research, simple seems to be better than complex. The constant correlations model seems to work best, followed by the Sharpe single-index model, as a forecaster of future correlation.

---

[14]Salomon Brothers argues that year-to-year changes are better than shorter time interval fluctuations because shorter sampling intervals result in greater volatility and therefore do not provide a reliable indicator of economic growth.

[15]Based on the generally accepted premise that the current default-free rate of interest (on Treasury bills) is composed of the cost of credit when inflation is zero plus the expected rate of inflation, Salomon extracts an expected inflation series from returns on Treasury bills using econometric methods.

It seems that more complex models tend to pick up more random noise than information. But the question remains: is it possible to combine any of these techniques with a second technique and improve estimation accuracy? Ledoit and Wolf (2003) derive optimum rules (shrinkage procedure) for combining forecasts from two different methods into a single forecast. They test optimum combinations of alternative models using Ledoit's shrinkage procedures and determine that a combination of historic pairwise values and forecasts from the Sharpe single-index works best. However, in 2004, Ledoit and Wolf found that combining historic pairwise values with the constant correlation model works even better.

In yet another round of this analysis, Elton, Gruber, and Spitzer (2006) find that better forecasts can be prepared by a two-step procedure. Step 1 involves forecasting the future level of the average correlation among stocks, while step 2 involves forecasting future difference from the mean. They find that an exponential smooth (with a smoothing coefficient of 0.5) or a rolling average of a past series of average correlation coefficients works best in predicting the future average correlation coefficient. They also find that breaking the overall population into groups based on industry membership or firm characteristics (size and beta) and assuming that the average correlation within each group, and among stocks in any two particular pairs of groups, is the same and equal to its historical value improves forecasting results.

CONCLUSION

In this chapter we have discussed alternatives to the single-index model for predicting future correlation coefficients. There are an infinite number of such models. Thus we cannot give definitive answers concerning their performance relative to single-index models. Many of the results are promising. This probably does not surprise the reader. What surprises most students is the ability of simple models, such as the single-index model and overall mean, to outperform more complex models in many tests. Although complex models better describe the historical correlation, they often contain more noise than information with respect to prediction. There is still a great deal of work to be done before complicated models consistently outperform simpler ones.

APPENDIX A

PROCEDURE FOR REDUCING ANY MULTI-INDEX MODEL TO A MULTI-INDEX MODEL WITH ORTHOGONAL INDEXES

We illustrate the procedure with a two-index model. Let

$$R_i = a_i^* + b_{i1}^* I_1^* + b_{i2}^* I_2^* + c_i$$

For example, I_1^* might be a market index and I_2^* a sector index (e.g., aggregate index for companies producing capital goods). If two indexes are correlated, the correlation may be removed from either index.

Define I_1 as equal to I_1^*. Now to remove the impact of the market from the sector index, we can establish the parameters of the following equation via regression analysis:

$$I_2^* = \gamma_0 + \gamma_1 I_1 + d_t$$

where γ_0 and γ_1 are regression coefficients and d_t is the random error term. By the techniques of estimation used in regression analysis, d_t is uncorrelated with I_1. Thus

$$d_t = I_2^* - (\gamma_0 + \gamma_1 I_1)$$

is an index of the performance of the sector index with the effect of I_1 (the market) removed.[16]

If we define

$$I_2 = d_t = I_2^* - \gamma_0 - \gamma_1 I_1$$

we have defined an index of sector performance that is uncorrelated with the market. Solving for I_2^* and substituting into the return equation yields

$$R_i = a_i^* + b_{i1}^* I_1 + b_{i2}^* I_2 + b_{i2}^* \gamma_0 + b_{i2}^* \gamma_1 I_1 + c_i$$

Rearranging terms gives

$$R_i = \left(a_i^* + b_{i2}^* \gamma_0\right) + \left(b_{i1}^* + b_{i2}^* \gamma_1\right) I_1 + b_{i2}^* I_2 + c_i$$

The first term is a constant we define as a_i. The coefficient on the second term is a constant we define as b_{i1}. Now let $b_{i2} = b_{i2}^*$. Then this equation becomes

$$R_i = a_i + b_{i1} I_1 + b_{i2} I_2 + c_i$$

where I_1 and I_2 have been defined so that they are uncorrelated, and we have accomplished our task.

If the model contained a third index, for example, an industry index, then this index could be made orthogonal to the other two indexes by running the following regression:

$$I_3^* = \theta_1 + \theta_2 I_2 + \theta_3 I_2 + e_i$$

The index I_3 could be defined as

$$I_3 = I_3^* - \left(\theta_1 + \theta_2 I_1 + \theta_3 I_2\right)$$

The proof that this leads to a three-index model with uncorrelated indexes of the form

$$R_i = a_i + b_{i1} I_1 + b_{i2} I_2 + b_{i3} I_3 + c_i$$

is left as an exercise to the reader.

APPENDIX B

MEAN RETURN, VARIANCE, AND COVARIANCE OF A MULTI-INDEX MODEL

In this appendix we derive the mean return variance and covariance of return when the multi-index model is assumed to describe the return structure in the market.

Expected Return

The expected return on a security with the multi-index model is

$$E\left(R_i\right) = E\left(a_i + b_{i1} I_1 + b_{i2} I_2 + \cdots + b_{iL} I_L + c_i\right)$$

Because the expected value of the sum of random variables is the sum of the expected values, we have

$$E\left(R_i\right) = E\left(a_i\right) + E\left(b_{i1} I_1\right) + E\left(b_{i2} I_2\right) + \cdots + E\left(b_{iL} I_L\right) + E\left(c_i\right)$$

[16]The index could also be written as $d_t + \gamma_0$. Each is appropriate. The way we have defined it, the mean is zero.

Recognizing that a and b are constants and that, by construction, $E(c_i) = 0$, we have

$$E(R_i) = a_i + b_{i1}\bar{I}_1 + b_{i2}\bar{I}_2 + \cdots + b_{iL}\bar{I}_L$$

This is the result stated in the text.

Variance of Return

The variance of the return on a security is

$$\sigma_i^2 = E\left(R_i - \bar{R}_i\right)^2$$

Substituting for R_i and $\bar{R}_i$, we have

$$\sigma_i^2 = E\Big[\left(a_i + b_{i1}I_1 + b_{i2}I_2 + \cdots + b_{iL}I_L + c_i\right) - \left(a_i + b_{i1}\bar{I}_1 + b_{i2}\bar{I}_2 + \cdots + b_{iL}\bar{I}_L\right)\Big]^2$$

Canceling the a_is and rearranging yields

$$\sigma_i^2 = E\Big[b_{i1}\left(I_1 - \bar{I}_1\right) + b_{i2}\left(I_2 - \bar{I}_2\right) + \cdots + b_{iL}\left(I_L - \bar{I}_L\right) + c_i\Big]^2$$

The next step is to square the terms in the brackets. The results of this can be seen if we examine all terms involving the first index. The first index times itself and each of the other terms is

$$E\Big[b_{i1}^2\left(I_1 - \bar{I}_1\right)^2 + b_{i1}b_{i2}\left(I_1 - \bar{I}_1\right)\left(I_2 - \bar{I}_2\right) + \cdots + b_{i1}b_{iL}\left(I_1 - \bar{I}_1\right)\left(I_L - \bar{I}_L\right) + b_{i1}\left(I_1 - \bar{I}_1\right)(c_i)\Big]$$

The expected value of the sum of random variables is the sum of the expected values, and since the b_is are constants, we have

$$b_{i1}^2 E\left(I_1 - \bar{I}_1\right)^2 + b_{i1}b_{i2}E\left[\left(I_1 - \bar{I}_1\right)\left(I_2 - \bar{I}_2\right)\right] + \cdots + b_{i1}b_{iL}E\left[\left(I_1 - \bar{I}_1\right)\left(I_L - \bar{I}_L\right)\right] + b_{i1}E\left[\left(I_1 - \bar{I}_1\right)(c_i)\right]$$

By construction

$$E\left[\left(I_i - \bar{I}_i\right)\left(I_j - \bar{I}_j\right)\right] = 0$$

and

$$E\left[\left(I_1 - \bar{I}_1\right)(c_i)\right] = 0$$

thus the only nonzero term involving index 1 is

$$b_{i1}^2 E\left(I_1 - \bar{I}_1\right)^2 = b_{i1}^2\sigma_{I1}^2$$

When we examine terms involving the c_i, we get the c_i with each index that has an expected value of zero. We also get $E(c_i)^2 = \sigma_{ci}^2$; thus

$$\sigma_i^2 = b_{i1}^2\sigma_{I1}^2 + b_{i2}^2\sigma_{I2}^2 + \cdots + b_{iL}^2\sigma_{IL}^2 + \sigma_{ci}^2$$

The Covariance

The covariance between securities i and j is

$$\sigma_{ij} = E\left[\left(R_i - \bar{R}_i\right)\left(R_j - \bar{R}_j\right)\right]$$

Substituting in the expressions for R_i and R_j yields

$$\sigma_{ij} = E\left\{\left[\left(a_i + b_{i1}I_1 + b_{i2}I_2 + \cdots + b_{iL}I_L + C_i\right) - \left(a_i + b_{i1}\bar{I}_1 + b_{i2}\bar{I}_2 + \cdots + b_{iL}\bar{I}_L\right)\right] \cdot \left[\left(a_j + b_{j1}I_1 + b_{j2}I_2 + \cdots + b_{jL}I_L + C_j\right) - \left(a_j + b_{j1}\bar{I}_1 + b_{j2}\bar{I}_2 + \cdots + b_{jL}\bar{I}_L\right)\right]\right\}$$

Noting that the *a*s cancel, and combining the terms involving the same *b*s, yields

$$\sigma_{ij} = E\left\{\left[b_{i1}\left(I_1 - \bar{I}_1\right) + b_{i2}\left(I_2 - \bar{I}_2\right) + \cdots + b_{iL}\left(I_L - \bar{I}_L\right) + c_i\right] \cdot \left[b_{j1}\left(I_1 - \bar{I}_1\right) + b_{j2}\left(I_2 - \bar{I}_2\right) + \cdots + b_{jL}\left(I_L - \bar{I}_L\right) + c_j\right]\right\}$$

The next step is to multiply out the terms. The results of this multiplication can be seen by considering the terms involving b_{i1}. They are

$$E\left[b_{i1}b_{j1}\left(I_1 - \bar{I}_1\right)^2 + b_{i1}b_{j2}\left(I_1 - \bar{I}_1\right)\left(I_2 - \bar{I}_2\right) + b_{i1}b_{j3}\left(I_1 - \bar{I}_1\right)\left(I_3 - \bar{I}_3\right) + \cdots + b_{i1}b_{jL}\left(I_1 - \bar{I}_1\right)\left(I_L - \bar{I}_L\right) + b_{i1}\left(I_1 - \bar{I}_1\right)c_j\right]$$

The expected value of all terms involving different indexes, for example, $(I_1 - \bar{I}_1)(I_k - \bar{I}_k)$, is zero by construction. Furthermore, the expected value of $b_{i1}(I_1 - \bar{I}_1)c_j$ is zero by construction. Thus the only nonzero term is

$$b_{i1}b_{j1}E\left(I_1 - \bar{I}_1\right)^2 = b_{i1}b_{j1}\sigma_{I1}^2$$

There are two types of terms involving the *c*s. First, there are terms like $b_{ik}(I_k - \bar{I}_k)c_j$, which is zero by construction. Second, there is the term c_ic_j. This is zero by assumption. Thus

$$\sigma_{ij} = b_{i1}b_{j1}\sigma_{I1}^2 + b_{i2}b_{j2}\sigma_{I2}^2 + b_{i3}b_{j3}\sigma_{I3}^2 + \cdots + b_{iL}b_{jL}\sigma_{IL}^2$$

QUESTIONS AND PROBLEMS

1. Given that the correlation coefficient between all securities is the same, call it ρ^*, and the assumption of the single-index model is accepted, derive an expression for the beta on any stock in terms of ρ^*.

2. Complete the procedure in Appendix A for reducing a general three-index model to a three-index model with orthogonal indexes.

3. Assume that all assumptions of the single-index model hold, except that the covariance between residuals is a constant K instead of zero. Derive the covariance between the two securities and the variance on a portfolio.

4. Given a three-index model such that all indexes are orthogonal, derive the formulas for the expected return, variance, and covariance of any stock.

5.

	Security		
	A	B	C
a_i	2	3	1
b_{i1}	0.8	1.1	0.9
b_{i2}	0.9	1.3	1.1
σ_{ci}	2.0	1.0	1.5

Assuming Is are uncorrelated and $\bar{I}_1 = 8$, $\bar{I}_2 = 4$, $\sigma_{I1} = 2.0$, $\sigma_{I2} = 2.5$, calculate the following using the general multi-index model:

(1) Expected returns

(2) Variance of return

(3) Covariance of return

6. Using the data from Problem 5, assume the model is now an Industry Index Model where $I_1 = I_m$ and that I_2 is now an industry index. Assuming that firms A and B are in the same industry, calculate the covariance of returns.

7. Repeat Problem 6, assuming now that firms B and C are in the same industry.

8. Given the multi-index model

$$R_i = 2 + 1.1I_1^* + 1.2I_2^* + C_i$$

where I_1^* and I_2^* are correlated, and given the regression equation $I_2^* = 1 + 1.3I_1 + d_t$, transform the equation for R_i into one with orthogonal indexes.

BIBLIOGRAPHY

1. Aber, John. "Industry Effects and Multivariate Stock Price Behavior," *Journal of Financial and Quantitative Analysis*, **XI,** No. 4 (Nov. 1976), pp. 617–624.
2. Bell, Frederick. "The Relation of the Structure of Common Stock Prices to Historical, Exceptional and Industrial Variables," *Journal of Finance*, **IX,** No. 1 (March 1976), pp. 187–197.
3. Berry, Michael, Brumeister, Edwin, and McElroy, Marjorie. "Sorting Out Risks Using Known APT Factors," *Financial Analysts Journal* **44,** (March 1988), pp. 29–42.
4. Brown, Stephen J. "The Number of Factors in Security Returns," *Journal of Finance*, **44,** No. 5 (1989), pp. 1247–1262.
5. Burmeister, Edwin, and McElroy, Marjorie. "APT and Multifactor Asset Pricing Models with Measured and Unobserved Factors: Theoretical and Econometric Issues," Discussion Paper, Department of Economics, University of Virginia and Duke University (1987).
6. ——. "Joint Estimation of Factor Sensitivities and Risk Premia for the Arbitrage Pricing Theory," *Journal of Finance*, **43,** No. 3 (July 1988), pp. 721–733.
7. Burmeister, Edwin, and Wall, Kent. "The Arbitrage Pricing Theory and Macroeconomic Factor Measures," *Financial Review* (Feb. 1986), pp. 1–20.
8. Burmeister, Edwin, Wall, Kent, and Hamilton, James. "Estimation of Unobserved Expected Monthly Inflation Using Kalman Filtering," *Journal of Business and Economic Statistics*, **4** (April 1986), pp. 147–160.
9. Carhart, Mark. "On Persistence in Mutual Fund Performance," *Journal of Finance*, **52,** (March 1997), pp. 661–692.
10. Campbell, John, and Shiller, Robert. "The Dividend Price Ratio and Expectation of Future Dividends and Discount Factors," *Review of Financial Studies*, **1** (1989), pp. 195–228.

11. Chan, Louis K. C., Karceski, Jason, and Lakonishok, Josef. "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model," *Review of Financial Studies*, **12** (1999), pp. 263–278.
12. Chen, Nai-fu. "Some Empirical Tests of the Theory of Arbitrage Pricing," *Journal of Finance*, **38** (Dec. 1983), pp. 1392–1414.
13. Chen, Nai-fu, Roll, Richard, and Ross, Stephen. "Economic Forces and the Stock Market," *Journal of Business*, **59** (July 1986), pp. 386–403.
14. Cho, D. Chinhyung, and Taylor, William. "The Seasonal Stability of the Factor Structure of Stock Returns," *Journal of Business*, **42** (Dec. 1987), pp. 1195–1211.
15. Cohen, Kalman, and Pogue, Jerry. "An Empirical Evaluation of Alternative Portfolio Selection Models," *Journal of Business*, **46** (April 1967), pp. 166–193.
16. Connor, G., and Korajczyk, R. "A Test for the Number of Factors in an Approximate Factor Model," *Journal of Finance*, **48** (1993), pp. 1263–1291.
17. Dhrymes, Phoebus, Friend, Irwin, and Gultekin, Bulent. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory," *The Journal of Finance*, **39** (June 1984), pp. 323–346.
18. Elton, Edwin J., and Gruber, Martin J. "Homogeneous Groups and the Testing of Economic Hypotheses," *Journal of Financial and Quantitative Analysis*, **IV**, No. 5 (Jan. 1970), pp. 581–602.
19. ——. "Improved Forecasting through the Design of Homogenous Groups," *Journal of Business*, **44,** No. 4 (Oct. 1971) pp. 432–450.
20. ——. "Estimating the Dependence Structure of Share Prices—Implications for Portfolio Selection," *Journal of Finance*, **VIII,** No. 5 (Dec. 1973), pp. 1203–1232.
21. Elton, Edwin J., Gruber, Martin J., and Spitzer, Jonathan. "Improved Estimates of Correlation Coefficients and Their Impact on the Optimum Portfolios," *European Financial Management*, **12,** No. 3 (June 2006), pp. 303–318.
22. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Common Factors in Fund Returns," *European Finance Review*, **3,** No. 4 (1999), pp. 320–332.
23. Elton, Edwin J., Gruber, Martin J., and Urich, Thomas. "Are Betas Best?" *Journal of Finance*, **23,** No. 5 (Dec. 1978), pp. 1375–1384.
24. Eun, Cheol S., and Resnick, Bruce G. "Estimating the Correlation Structure of International Share Prices," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1311–1324.
25. Fama, Eugene. "Stock Returns, Real Activity, Inflation and Money," *American Economic Review*, **71** (1981), pp. 545–565.
26. Fama, Eugene, and Gibbons, Michael. "A Comparison of Inflation Forecasts," *Journal of Monetary Economics*, **13** (1984), pp. 327–348.
27. Fama, Eugene, and MacBeth, James. "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, **38** (1973), pp. 607–636.
28. Fama, Eugene, and French, Kenneth. "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, **33** (1993), pp. 3–56.
29. Farrell, James. "Analyzing Covariation of Returns to Determine Homogenous Stock Groupings," *Journal of Business*, **47,** No. 2 (April 1974), pp. 186–207.
30. ——. *The Multi-Index Model and Practical Portfolio Analysis*, The Financial Analysts Research Foundation Occasional Paper No. 4 (1976).
31. Fertuck, Leonard. "A Test of Industry Indexes Based on SIC Codes," *Journal of Financial and Quantitative Analysis*, **X,** No. 5 (Dec. 1975), pp. 837–848.
32. Gibbons, Michael R. "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics*, **10** (March 1982), pp. 3–27.
33. Grinblatt, Mark, and Titman, Sheridan. "Approximate Factor Structures: Interpretations and Implications for Empirical Tests," *Journal of Finance*, **40** (1985), pp. 1367–1373.
34. Gultekin, Mustafa, and Gultekin, N. Bulent. "Stock Return Anomalies and Tests of the APT," *Journal of Finance*, **42** (Dec. 1987), pp. 1213–1224.
35. Hansen, Lars, and Singleton, Kenneth. "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Assets Returns," *Journal of Political Economy*, **91** (1983), pp. 249–265.
36. Huberman, Gur, and Kandel, Shmuel. "Mean–Variance Spanning," *Journal of Finance*, **42** (Sept. 1987), pp. 873–888.

37. Ibbotson, Roger, and Sinquefield, Rex. *Stocks, Bonds, Bills and Inflation: The Past and the Future* (Charlottesville, VA: Financial Analysts Research Foundation, 1982).
38. Jagganthan, Ravi, and Ma, Tongshu. "Risk Reduction in Large Portfolios: Why Imposing Wrong Constraints Helps," *Journal of Finance*, **58** (2003), pp. 1651–1684.
39. King, Benjamin. "Market and Industry Factors in Stock Price Behavior," *Journal of Business*, **39** (Jan. 1966), pp. 139–140.
40. Kryzanowski, Lawrence, and To, Minh Chan. "General Factor Models and the Structure of Security Returns," *Journal of Financial and Quantitative Analysis*, **18** (1983), pp. 31–52.
41. Ledoit, Olivier, and Wolf, Michael. "Improved Estimation of the Covariance Matrix of Stock Returns with an Application to Portfolio Selection," *Journal of Empirical Finance*, **10,** No. 5 (Dec. 2003) pp. 603–624.
42. Ledoit, Olivier, and Wolf, Michael. "Honey, I Shrunk the Sample Covariance Matrix," *Journal of Portfolio Management*, **31,** No. 1 (2004) pp. 1–20.
43. Martin, John, and Klemkosky, Robert. "The Effect of Homogeneous Stock Groupings on Portfolio Risk," *Journal of Business*, **49,** No. 3 (July 1976), pp. 339–349.
44. McElroy, Marjorie, and Burmeister, Edwin. "Arbitrage Pricing Theory as a Restricted Nonlinear Multivariate Regresssion Model: ITNLSUR Estimates," *Journal of Business and Economic Statistics*, **VI,** No. 1 (Jan. 1988), pp. 29–42.
45. McElroy, Marjorie, and Wall, Kent. "Two Estimators for the APT Model When Factors Are Measured," *Economics Letters*, **19** (1985), pp. 271–275.
46. Merton, Robert C. "An Intertemporal Capital Asset Pricing Model," *Econometrica*, **41** (1973), pp. 867–887.
47. Meyers, Stephen. "A Re-examination of Market and Industry Factors in Stock Price Behavior," *Journal of Finance*, **VIII,** No. 3 (June 1973), pp. 695–705.
48. Morgan, I. G. "Grouping Procedures for Portfolio Formation," *Journal of Finance*, **XI,** No. 5 (Dec. 1977), pp. 1759–1765.
49. Ohlson, James, and Garman, Mark. "A Dynamic Equilibrium for the Ross Arbitrage Model," *Journal of Finance*, **35** (1980), pp. 675–684.
50. Reilly, Frank, and Dryzycimski, Eugene. "Alternative Industry Performance and Risk," *Journal of Financial and Quantitative Analysis*, **IX,** No. 3 (June 1974), pp. 423–446.
51. Roll, Richard, and Ross, Stephen. "An Empirical Investigation of the Arbitrage Pricing Theory," *The Journal of Finance*, **35** (Dec. 1980), pp. 1073–1103.
52. Rosenberg, Barr. "Extra-Market Components of Covariance in Security Returns," *Journal of Financial and Quantitative Analysis*, **IX,** No. 2 (March 1974), pp. 263–274.
53. Sorensen, Eric, Salomon, R. S., Davenport, Caroline, and Fiore, Maria. *Risk Analysis: The Effect of Key Macroeconomic and Market Factors on Portfolio Returns* (New York: Salomon Brothers, 1989).

9

Simple Techniques for Determining the Efficient Frontier

In Chapters 7 and 8 we examined several models that were developed to simplify the inputs to the portfolio selection problem. Each of these models makes an assumption about why stocks covary together. Each leads to a simplified structure for the correlation matrix or covariance matrix between securities. These models were developed to cut down on the number of inputs and simplify the nature of the inputs needed to forecast correlations between securities. The use of these models was expected to lead to some loss of accuracy in forecasting correlations, but the ease of using the models was expected to compensate for this loss of accuracy. However, we have seen in Chapters 7 and 8 that when fitted to historical data, these simplifying models result in an increase, not a decrease, in forecasting accuracy. The models are of major interest because they both reduce and simplify the inputs needed to perform portfolio analysis *and* increase the accuracy with which correlations and covariances can be forecast.

In this chapter we see that there is yet another advantage to these models. Each allows the development of a system for computing the composition of optimum portfolios that is so simple it can often be performed without the use of a computer. Perhaps even more important than the ease of computation is the fact that the methods of portfolio selection described in this chapter make it very clear why a stock does or does not enter into an optimal portfolio. Each model of the correlation structure discussed in Chapters 7 and 8 leads to a unique ranking of stocks, such that if a stock enters an optimal portfolio, any higher-ranked stock must also enter the optimal portfolio. Similarly, if a stock does not enter an optimal portfolio, any lower-ranked stock does not enter the optimal portfolio. This allows the analyst to judge the relative desirability of stocks even before the portfolio selection process is begun. Furthermore, as we will see, the optimum ranking of stocks depends on variables that are already familiar to security analysts and portfolio managers, as well as to readers of this book. This should minimize the institutional barriers to their adoption.

In this chapter we describe, in detail, the methods for selecting optimal portfolios that are appropriate when the single-index model and the constant-correlation model are accepted as descriptions of the covariance structure between securities. In the text of this chapter we present the rules for optimal portfolio selection and show how to use them. This may appear as magic to the reader because, while we declare that the rules lead to the

selection of optimal portfolios, the text does not contain a proof that this is so. For the reader who prefers science to magic, the appendices at the end of this chapter present the derivations of all of the rules described in the text. These derivations also act as proof of the optimality of the rules. We have separated the material in this way because the mathematical sophistication needed to understand the derivation of the rules is so much greater than the mathematical sophistication needed to use the rules.

We close this chapter with a brief discussion of the types of rules to which some of the other models of correlation structure (presented in Chapter 8) lead. The discussion here is quite concise, but, for the reader interested in learning more about these rules, the appropriate references are noted.

THE SINGLE-INDEX MODEL

In this section we present and demonstrate the optimum procedure for selecting portfolios when the single-index model is accepted as the best way to forecast the covariance structure of returns.

First we present the ranking criteria that can be used to order stocks for selection for the optimal portfolio. We next present the technique for employing this ranking device to form an optimum portfolio, along with a logical explanation for why it works. While the technique for forming optimum portfolios is easy to understand, the formal proof that it leads to the same portfolio that would be produced by the optimum procedure, presented in Chapter 6, is complex and is presented in Appendix A and Appendix C at the end of this chapter.

After presenting the criteria for the composition of an optimal portfolio, we demonstrate its use with some simple examples. In the first part of the section we assume that short sales are forbidden. In the latter part we allow short sales. In addition, we start by assuming unlimited borrowing and lending at the riskless rate. This assumption is dropped later in the chapter.

The Formation of Optimal Portfolios

The calculation of optimal portfolios would be greatly facilitated, and the ability of practicing security analysts and portfolio managers to relate to the construction of optimal portfolios greatly enhanced, if there were a single number that measured the desirability of including a stock in the optimal portfolio. If one is willing to accept the standard form of the single-index model as describing the comovement between securities, such a number exists. In this case, the desirability of any stock is directly related to its excess return to beta ratio. Excess return is the difference between the expected return on the stock and the riskless rate of interest such as the rate on a Treasury bill. The excess return to beta ratio measures the additional return on a security (beyond that offered by a riskless asset) per unit of nondiversifiable risk. The form of this ratio should lead to its easy interpretation and acceptance by security analysts and portfolio managers because they are used to thinking in terms of the relationship between potential rewards and risk.[1] The numerator of this ranking device is the extra return over the riskless asset that we earn from holding a security

[1]In Chapter 18 we see that one commonly used measure to rank portfolio performance is the portfolio's excess return to beta ratio, with the best portfolio being the one with the highest ratio. It is intuitively appealing to rank stocks by the same criteria as one uses to rank portfolios, and in fact, it is shown in Appendix A that it is optimal to do so.

other than the riskless asset. The denominator is the nondiversifiable risk (the risk we cannot get rid of) that we are subject to by holding a risky security rather than the riskless asset.

More formally, the index we use to rank stocks is *excess return to beta*, or

$$\frac{\bar{R}_i - R_F}{\beta_i}$$

where

$\bar{R}_i =$ the expected return on stock i

$R_F =$ the return on a riskless asset

$\beta_i =$ the expected change in the rate of return on stock i associated with a 1% change in the market return

If stocks are ranked by excess return to beta (from highest to lowest), the ranking represents the desirability of any stock's inclusion in a portfolio. In other words, if a stock with a particular ratio of $(\bar{R}_i - R_F)/\beta_i$ is included in an optimal portfolio, all stocks with a higher ratio will also be included. Conversely, if a stock with a particular $(\bar{R}_i - R_F)/\beta_i$ is excluded from an optimal portfolio, all stocks with lower ratios will be excluded (or if short selling is allowed, sold short). When the single-index model is assumed to represent the covariance structure of security returns, then a stock is included or excluded depending only on the size of its excess return to beta ratio. How many stocks are selected depends on a unique cutoff rate such that all stocks with higher ratios of $(\bar{R}_i - R_F)/\beta_i$ will be included and all stocks with lower ratios will be excluded. We call this cutoff ratio C^*.

The rules for determining which stocks are included in the optimum portfolio are as follows:

1. Find the excess return to beta ratio for each stock under consideration and rank from highest to lowest.
2. The optimum portfolio consists of investing in all stocks for which $(\bar{R}_i - R_F)/\beta_i$ is greater than a particular cutoff point C^*. Shortly, we will define C^* and interpret its economic significance.

The preceding procedure is extremely simple. Once C^* has been determined, the securities to be included can be selected by inspection. Furthermore, the amount to invest in each security is equally simple to determine, as is discussed shortly.

Ranking Securities

In Tables 9.1 and 9.2 we present an example that illustrates this procedure. Table 9.1 contains the data necessary to apply our simple ranking device to determine an optimal portfolio. It is the normal output generated from a single-index or beta model, plus the ratio of excess return to beta. These same data could alternatively be generated by analysts' subjective estimates. There are 10 securities in the tables. For the reader's convenience, we have already ranked the securities according to $(\bar{R}_i - R_F)/\beta_i$ and have used numbers that make the calculations easy to follow. The application of rule 2 involves the comparison of $(\bar{R}_i - R_F)/\beta_i$ with C^*. Accept that $C^* = 5.45$ for the moment; we will shortly present a procedure for its calculation. Examining Table 9.1 shows that for securities 1 to 5, $(\bar{R}_i - R_F)/\beta_i$ is greater than C^*, while for security 6, it is less than C^*. Hence, an optimal portfolio consists of securities 1 to 5.

Table 9.1 Data Required to Determine Optimal Portfolio $R_F = 5\%$

1	2	3	4	5	6
Security No. i	Mean Return $\bar{R}_i$	Excess Return $\bar{R}_i - R_F$	Beta β_i	Unsystematic Risk σ_{ei}^2	Excess Return over Beta $\frac{(\bar{R}_i - R_F)}{\beta_i}$
1	15	10	1	50	10
2	17	12	1.5	40	8
3	12	7	1	20	7
4	17	12	2	10	6
5	11	6	1	40	6
6	11	6	1.5	30	4
7	11	6	2	40	3
8	7	2	0.8	16	2.5
9	7	2	1	20	2
10	5.6	0.6	0.6	6	1.0

Table 9.2 Calculations for Determining Cutoff Rate with $\sigma_m^2 = 10$

1	2	3	4	5	6	7
Security No. i	$\frac{(\bar{R}_i - R_F)}{\beta_i}$	$\frac{(\bar{R}_i - R_F)\beta_i}{\sigma_{ei}^2}$	$\frac{\beta_i^2}{\sigma_{ei}^2}$	$\sum_{j=1}^{i} \frac{(\bar{R}_j - R_F)\beta_j}{\sigma_{ej}^2}$	$\sum_{j=1}^{i} \frac{\beta_j^2}{\sigma_{ej}^2}$	C_i
1	10	2/10	2/100	2/10	2/100	1.67
2	8	4.5/10	5.625/100	6.5/10	7.625/100	3.69
3	7	3.5/10	5/100	10/10	12.625/100	4.42
4	6	24/10	40/100	34/10	52.625/100	5.43
5	6	1.5/10	2.5/100	35.5/10	55.125/100	5.45
6	4	3/10	7.5/100	38.5/10	62.625/100	5.30
7	3	3/10	10/100	41.5/10	72.625/100	5.02
8	2.5	1/10	4/100	42.5/10	76.625/100	4.91
9	2.0	1/10	5/100	43.5/10	81.625/100	4.75
10	1.0	0.6/10	6/100	44.1/10	87.625/100	4.52

Setting the Cutoff Rate (C^*)

As discussed earlier, C^* is the cutoff rate. All securities whose excess return to risk ratio is above the cutoff rate are selected, and all whose ratios are below are rejected. The value of C^* is computed from the characteristics of all of the securities that belong in the optimum portfolio. To determine C^*, it is necessary to calculate its value as if there were different numbers of securities in the optimum portfolio. Designate C_i as a candidate for C^*. The value of C_i is calculated when i securities are assumed to belong to the optimal portfolio.

Because securities are ranked from highest excess return to beta to lowest, we know that if a particular security belongs in the optimal portfolio, all higher-ranked securities also belong in the optimal portfolio. We proceed to calculate values of a variable C_i (the procedure is outlined below) as if the first ranked security were in the optimal portfolio ($i = 1$), then the first- and second-ranked securities were in the optimal portfolio ($i = 2$), then the first-, second-, and third-ranked securities were in the optimal portfolio ($i = 3$), and so forth. These C_i are candidates for C^*. We know we have found the optimum C_i—that is, C^*—when all securities used in the calculation of C_i have excess returns to beta above C_i and all

securities not used to calculate C_i have excess returns to beta below C_i. For example, column 7 of Table 9.2 shows the C_i for alternative values of *i*. Examining the table shows that C_5 is the only value of C_i for which all securities used in the calculation of *i* (1 through 5 in the table) have a ratio of excess return to beta above C_i and all securities not used in the calculation of C_i (6 through 10 in the table) have an excess return to beta ratio below C_i. C_5 serves the role of a cutoff rate in the way a cutoff rate was defined earlier. In particular, C_5 is the only C_i that, when used as a cutoff rate, selects only the stocks used to construct it. There will always be one and only one C_i with this property, and it is C^*.

Calculating the Cutoff Rate C*

Recall that stocks are ranked by excess return to risk from highest to lowest. For a portfolio of *i* stocks, C_i is given by

$$C_i = \frac{\sigma_m^2 \sum_{j=1}^{i} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \sum_{j=1}^{i} \left(\frac{\beta_j^2}{\sigma_{ej}^2} \right)} \tag{9.1}$$

where

σ_m^2 = the variance in the market index

σ_{ej}^2 = the variance of a stock's movement that is not associated with the movement of the market index; this is usually referred to as a stock's unsystematic risk

This looks daunting. But a moment's reflection combined with a peek at the example below will show that it is not as hard to compute as it appears. Although Equation (9.1) is the form that should actually be used to compute C_i, this expression can be stated in a mathematically equivalent way that clarifies the meaning of C_i.[2]:

$$C_i = \frac{\beta_{iP}\left(\overline{R}_P - R_F\right)}{\beta_i} \tag{9.2}$$

where

β_{iP} = the expected change in the rate of return on stock *i* associated with a 1% change in the return on the optimal portfolio

$\overline{R}_P$ = the expected return on the optimal portfolio

All other terms as before.

Variables β_{iP} and $\overline{R}_P$ are, of course, not known until the optimal portfolio is determined. Hence Equation (9.2) could not be used to actually determine the optimum portfolio; rather, Equation (9.1) must be used. However, this expression for C_i is useful in interpreting the economic significance of our procedure. Recall that securities are added to the portfolio as long as

$$\frac{\overline{R}_i - R_F}{\beta_i} > C_i$$

[2]See Appendix A at the end of this chapter for a derivation of this expression.

Rearranging and substituting in Equation (9.2) yields

$$\left(\overline{R}_i - R_F\right) > \beta_{iP}\left(\overline{R}_P - R_F\right)$$

The right-hand side is nothing more than the expected excess return on a particular stock based solely on the expected performance of the optimum portfolio. The term on the left-hand side is the security analyst's estimate of the expected excess return on the individual stock. Thus, if the analysis of a particular stock leads the portfolio manager to believe that it will perform better than would be expected, based on its relationship to the optimal portfolio, it should be added to the portfolio.

Now let us look at how Equation (9.1) can be used to determine the value of C_i for our example. Although Equation (9.1) might look complex, the ease with which it can be calculated is demonstrated by Table 9.2. This table presents the intermediate calculations necessary to determine Equation (9.1).

Let us work through the intermediate calculations shown in Table 9.2 and find the value for C_i for the first security in our list of securities. The numerator of Equation (9.1) is

$$\sigma_m^2 \sum_{j=1}^{i} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}$$

Column 3 of Table 9.2 presents the value of

$$\frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}$$

for each security. This is necessary to determine the summation. For example, for the first security using the values shown in Table 9.1, it is

$$\frac{(15-5)1}{50} = \frac{2}{10}$$

Column 5 gives the value of the summation, or the running cumulative total of column 3. For the first security, $i = 1$ and

$$\sum_{j=1}^{1} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2} = \frac{\left(\overline{R}_1 - R_F\right)\beta_1}{\sigma_{e1}^2}$$

Thus column 5 of Table 9.2 is the same as column 3 for security 1. The last term in the denominator of Equation (9.1) is

$$\sum_{j=1}^{i} \frac{\beta_j^2}{\sigma_{ej}^2}$$

Since $i = 1$ for the first security, it is simply

$$\frac{\beta_1^2}{\sigma_{e1}^2} = \frac{(1)^2}{50} = \frac{2}{100}$$

This result is shown in column 4 and cumulated in column 6. We can now put these terms together to find C_i. Remembering that $\sigma_m^2 = 10$,

$$C_i = \frac{\sigma_m^2 \sum_{j=1}^{i} \frac{(\bar{R}_j - R_F)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \sum_{j=1}^{i} \frac{\beta_j^2}{\sigma_{ej}^2}}$$

$$= \frac{\sigma_m^2(\text{column 5})}{1 + \sigma_m^2(\text{column 6})} = \frac{10\left(\frac{2}{10}\right)}{1 + 10\left(\frac{2}{100}\right)} = 1.67$$

We now follow through the calculations for security 2 ($i = 2$). Column 3 is found to be

$$\frac{(17 - 5)1.5}{40} = \frac{4.5}{10}$$

Now column 5 is the sum of column 3 for security 1 and security 2, or

$$\frac{2}{10} + \frac{4.5}{10} = \frac{6.5}{10}$$

Column 4 is

$$\frac{(1.5)^2}{40} = \frac{5.625}{100}$$

Column 6 is the sum of column 4 for security 1 and 2, or

$$\frac{2}{100} + \frac{5.625}{100} = \frac{7.625}{100}$$

We can now find C_2 as

$$C_2 = \frac{\sigma_m^2(\text{column 5})}{1 + \sigma_m^2(\text{column 6})} = \frac{10\frac{6.5}{10}}{1 + 10\frac{7.625}{100}} = 3.68$$

Proceeding in the same fashion, we can find all the C_is.

Constructing the Optimal Portfolio

Once the securities that are contained in the optimum portfolio are determined, it remains to show how to calculate the percentage invested in each security. The percentage invested in each security is

$$X_i = \frac{Z_i}{\sum_{\text{included}} Z_j}$$

where

$$Z_i = \frac{\beta_i}{\sigma_{ei}^2}\left(\frac{\overline{R}_i - R_F}{\beta_i} - C^*\right) \tag{9.3}$$

The second expression determines the relative investment in each security, whereas the first expression simply scales the weights on each security so they sum to 1 and, thus, ensure full investment. Note that the residual variance on each security σ_{ei}^2 plays an important role in determining how much to invest in each security. Applying this formula to our example, we have

$$Z_1 = \frac{2}{100}(10 - 5.45) = 0.091$$
$$Z_2 = \frac{3.75}{100}(8 - 5.45) = 0.095625$$
$$Z_3 = \frac{5}{100}(7 - 5.45) = 0.0775$$
$$Z_4 = \frac{20}{100}(6 - 5.45) = 0.110$$
$$Z_5 = \frac{2.5}{100}(6 - 4.5) = 0.01375$$
$$\sum_{i=1}^{5} Z_i = 0.387875$$

Dividing each Z_i by the sum of the Z_i, we find that we should invest 23.5% of our funds in security 1, 24.6% in security 2, 20% in security 3, 28.4% in security 4, and 3.5% in security 5.

Let us stress that this is identical to the result that would have been achieved had the problem been solved using the established quadratic programming codes. However, the solution has been reached in a fraction of the time with a set of relatively simple calculations.

Notice that the characteristics of a stock that make it desirable and the relative attractiveness of stocks can be determined before the calculations of an optimal portfolio are begun. The desirability of any stock is solely a function of its excess return to beta ratio. Thus a security analyst following a set of stocks can determine the relative desirability of each stock before the information from all analysts is combined and the portfolio selection process begun.

Up to this point, we have assumed that all stocks have positive betas. We believe that there are sound economic reasons to expect all stocks to have positive betas and that the few negative beta stocks that are found in large samples are due to measurement errors. However, as pointed out in Elton, Gruber, and Padberg (1978), negative beta stocks (and zero beta stocks) are easily incorporated in the analysis.

Another Example

We have included a second example to illustrate the use of these formulas. This example is presented in Tables 9.3 and 9.4. Once again, securities are ranked by excess return to beta. Examining Table 9.4 shows that the C_i associated with security 4 is the only C_i

Table 9.3 Data Required to Determine Optimal Portfolio; $R_F = 5$

1	2	3	4	5	6
Security Number i	Mean Return $\bar{R}_i$	Excess Return $\bar{R}_i - R_F$	Beta β_i	Unsystematic Risk σ_{ei}^2	Excess Return over Beta $\frac{(\bar{R}_i - R_F)}{\beta_i}$
1	19	14	1.0	20	14
2	23	18	1.5	30	12
3	11	6	0.5	10	12
4	25	20	2.0	40	10
5	13	8	1.0	20	8
6	9	4	0.5	50	8
7	14	9	1.5	30	6
8	10	5	1.0	50	5
9	9.5	4.5	1.0	50	4.5
10	13	8	2.0	20	4
11	11	6	1.5	30	4
12	8	3	1.0	20	3
13	10	5	2.0	40	2.5
14	7	2	1.0	20	2

Table 9.4 Calculations for Determining Cutoff Rate with $\sigma_m^2 = 10$

Security Number i	$\frac{(\bar{R}_i - R_F)}{\beta_i}$	$\frac{(\bar{R}_i - R_F)\beta_i}{\sigma_{ei}^2}$	$\frac{\beta_i^2}{\sigma_{ei}^2}$	$\sum_{j=1}^{i} \frac{(\bar{R}_j - R_F)\beta_j}{\sigma_{ej}^2}$	$\sum_{j=1}^{i} \frac{\beta_j^2}{\sigma_{ej}^2}$	C_i
1	14	$\frac{70}{100}$	$\frac{5}{100}$	$\frac{70}{100}$	$\frac{5}{100}$	4.67
2	12	$\frac{90}{100}$	$\frac{7.5}{100}$	$\frac{160}{100}$	$\frac{12.5}{100}$	7.11
3	12	$\frac{30}{100}$	$\frac{2.5}{100}$	$\frac{190}{100}$	$\frac{15}{100}$	7.6
4	10	$\frac{100}{100}$	$\frac{10}{100}$	$\frac{290}{100}$	$\frac{25}{100}$	8.29
5	8	$\frac{40}{100}$	$\frac{5}{100}$	$\frac{330}{100}$	$\frac{30}{100}$	8.25
6	8	$\frac{4}{100}$	$\frac{0.5}{100}$	$\frac{334}{100}$	$\frac{30.5}{100}$	8.25
7	6	$\frac{45}{100}$	$\frac{7.5}{100}$	$\frac{379}{100}$	$\frac{38}{100}$	7.9
8	5	$\frac{10}{100}$	$\frac{2}{100}$	$\frac{389}{100}$	$\frac{40}{100}$	7.78
9	4.5	$\frac{9}{100}$	$\frac{2}{100}$	$\frac{398}{100}$	$\frac{42}{100}$	7.65
10	4	$\frac{80}{100}$	$\frac{20}{100}$	$\frac{478}{100}$	$\frac{62}{100}$	6.64
11	4	$\frac{30}{100}$	$\frac{7.5}{100}$	$\frac{508}{100}$	$\frac{69.5}{100}$	6.39
12	3	$\frac{15}{100}$	$\frac{5}{100}$	$\frac{523}{100}$	$\frac{74.5}{100}$	6.19
13	2.5	$\frac{25}{100}$	$\frac{10}{100}$	$\frac{548}{100}$	$\frac{84.5}{100}$	5.8
14	2	$\frac{10}{100}$	$\frac{5}{100}$	$\frac{558}{100}$	$\frac{89.5}{100}$	5.61

consistent with our definition of C^*. That is, it is the only value of C_i such that stocks ranked i or higher all have excess returns to beta above C_i and all stocks ranked below i have excess returns to beta below C_i, thus the cutoff rate

$$C^* = C_4 = \frac{58}{7} = 8.29$$

The optimum amount to invest is determined using Equation (9.3). For this example, it is

$$Z_1 = \frac{1}{20}\left(14 - \frac{58}{7}\right) = \frac{40}{140} = \frac{240}{840}$$
$$Z_2 = \frac{1.5}{30}\left(12 - \frac{58}{7}\right) = \frac{39}{210} = \frac{156}{840}$$
$$Z_3 = \frac{0.5}{10}\left(12 - \frac{58}{7}\right) = \frac{13}{70} = \frac{156}{840}$$
$$Z_4 = \frac{2}{40}\left(10 - \frac{58}{7}\right) = \frac{24}{280} = \frac{72}{840}$$

Scaling the Zs so that they add to 1, we have

$$X_1 = \frac{240}{240 + 156 + 156 + 72} = \frac{240}{624} = 0.38$$
$$X_2 = \frac{156}{240 + 156 + 156 + 72} = \frac{156}{624} = 0.25$$
$$X_3 = \frac{156}{240 + 156 + 156 + 72} = \frac{156}{624} = 0.25$$
$$X_4 = \frac{72}{240 + 156 + 156 + 72} = \frac{72}{624} = 0.12$$

Thus, in this example, the optimum portfolio consists of four securities, with the largest investment in security 1 and the smallest in security 4.

In solving this problem, there is no need to fill in all the entries in Table 9.4. Clearly all the intermediate calculations associated with the lower-ranked securities are not needed. One could start by ranking all securities by excess return to beta, and then proceed to calculate C_i for larger values of i (higher-ranked stocks) until a value of i is found so that the ith + 1 stock is excluded. At that point, we can ignore stocks ranked below the ith stock. Notice that, though excess return to beta had to be computed for all stocks, the calculation of C_i and Z_i need only be done for i stocks or, in the case of this example, four stocks.

Short Sales Allowed

The procedures used to calculate the optimal portfolio when short sales are allowed are closely related to the procedures in the no short sales case. As a first step, all stocks are ranked by excess return to beta, just as they were in the previous case. However, the cutoff point for stocks, C^*, now has a different meaning as well as a different procedure for calculation. When short sales are allowed, all stocks will either be held long or sold short.[3]

[3]Actually, it is possible for one or more stocks to have return and risk characteristics so that they are held in exactly zero proportions. This does not affect the procedure described in this section.

Thus all stocks enter into the optimum portfolio, and all stocks affect the cutoff point. Equation (9.1) still represents the cutoff point, but now the numerator and denominator of this equation are summed over all stocks. In addition, although Equations (9.1) and (9.3) still hold (with respect to the new C^*), the meaning of Z_i is now changed. We now have to calculate a value for Z_i for each stock. A positive value of Z_i indicates the stock will be held long, and a negative value indicates it will be sold short. Thus the impact of C^* has changed. Stocks that have an excess return to beta above C^* are held long (as before), but stocks with an excess return to beta below C^* are now sold short.

Let us illustrate this by returning to the first example presented earlier in Table 9.2. Remember, to calculate C^*, we must employ Equation (9.1) with i set equal to the number of stocks under consideration. In this case we have a population of 10 stocks so that

$$C^* = C_{10} = 4.52$$

Employing Equation (9.3) for each security, we find

$$Z_1 = \frac{1}{50}[10 - 4.52] = 0.110 \qquad Z_7 = \frac{2}{40}[3 - 4.52] = -0.076$$

$$Z_2 = \frac{1.5}{40}[8 - 4.52] = 0.131 \qquad Z_8 = \frac{0.8}{16}[2.5 - 4.52] = -0.101$$

$$Z_3 = \frac{1}{20}[7 - 4.52] = 0.124 \qquad Z_9 = \frac{1}{20}[2 - 4.52] = -0.126$$

$$Z_4 = \frac{2}{10}[6 - 4.52] = 0.296 \qquad Z_{10} = \frac{0.6}{6}[1.0 - 4.52] = -0.352$$

$$Z_5 = \frac{1}{40}[6 - 4.52] = 0.037$$

$$Z_6 = \frac{1.5}{30}[4 - 4.52] = -0.026 \qquad \sum_{i=1}^{10} Z_i = 0.017$$

The last step in the procedure involves the scaling of the Z_is so they represent the optimum proportions to invest in each stock (X_is). There are actually two ways to do this scaling. These methods exactly parallel the two definitions of short sales we examined in earlier chapters. Under the standard definition of short sales, which presumes that a short sale of a stock is a source of funds to the investor, the appropriate scaling factor is given by

$$X_i = \frac{Z_i}{\sum_{j=1}^{N} Z_j}$$

where Z_i can be positive or negative. This scaling factor is arrived at by realizing that under this definition of short sales, the constraint on the X_is is that

$$\sum_{i=1}^{N} X_i = 1$$

The second definition of short sales we referred to earlier is Lintner's definition. Under this definition, short sales are a use of the investor's funds; however, the investor receives

Table 9.5 Optimum Percentages

Security	Short Sales Disallowed	Lintner Definition of Short Sales	Standard Definition of Short Sales
1	23.5	8.0	647.1
2	24.6	9.5	770.6
3	20.0	9.0	729.4
4	28.4	21.5	1,741.2
5	3.5	2.7	217.6
6	0	−1.9	−152.9
7	0	−5.5	−447.1
8	0	−7.3	−594.1
9	0	−9.1	−741.2
10	0	−25.5	−2,070.6

the riskless rate of the funds involved in the short sale.[4] We have seen that this translates into the constraint

$$\sum_{i=1}^{N} |X_i| = 1$$

The analogous scaling factor is

$$X_i = \frac{Z_i}{\sum_{j=1}^{N} |Z_j|} \tag{9.4}$$

In Table 9.5 we have presented the fraction of funds that the investor should place in each security when short sales are not allowed, when the standard definition of short sales is employed, and when Lintner's definition of short sales is used.

Note that under the two alternative definitions of short sales, not only are the same stocks always held long and sold short but any two stocks are always held in the same ratio to each other. This is true because the two solutions differ by only a scale factor. From the foregoing analysis, it is obvious that this scale factor is simply

$$\frac{\sum_{i=1}^{N} |Z_i|}{\sum_{i=1}^{N} Z_i}$$

One point of interest this example makes clear is that employing the normal definition of short sales can really change the scale of the optimal solution. While the proportions invested under the Lintner definition seem reasonable, for example, place 8% of your money in security 1 and use 25.5% of your funds to short sell security 10, the solution that can be reached under the standard definition of short sales can seem extreme. In this example, the standard

[4]To be precise, the Lintner definition assumes that the proceeds of the short sale are not available for investment. Furthermore, the investor must put up an amount of funds equal to the proceeds of the short sale as collateral to protect against adverse price movements. The return on the short sale is the opposite of a long purchase. A negative value for X is required in determining the return on a portfolio. However, in analyzing the constraint on the amount invested, the additional funds invested must be considered—hence the absolute value sign in the constraint of the sum of Xs. See footnote 1 in Chapter 6 for a further explanation.

definition of short sales would involve investing in security 1 a sum of money equal to 6.47 times the amount originally available for investment and selling short an amount of security 10 equal to 20.7 times the amount originally available for investment.

If we now compare either of the short sales examples with the short sales disallowed examples, we can see some interesting differences. First, note that the proportion placed in any stock relative to a second stock need bear no relationship between the two cases. As an example, examine security 1 and security 4 in the short sales allowed and short sales not allowed examples. Both call for security 1 and security 4 to be held long. When short sales are not allowed, we hold 1.21 as much of security 4 as we hold of security 1. When short sales are allowed, we hold 2.69 as much of security 4 as we hold of security 1. This demonstrates that the proportions held of securities under short sales allowed need bear no particular relationship to the proportions held of the securities when short sales are not allowed.

In fact, although this particular example does not demonstrate it, the set of securities that is held long can be different according to whether short sales are allowed or not. This can be seen by reexamining example 2. When short sales were not allowed, we have seen that the first four securities are held long. If short sales are allowed, the appropriate value for C (all securities included) is 5.61 from Table 9.4. Examining Table 9.4, we now see that the first seven rather than the first four securities should be held long in the optimal portfolio.

The fact that allowing short sales changes the nature of the optimal solution should not come as a surprise to the reader. Allowing short sales is equivalent to adding new securities to the set from which the optimal portfolio will be selected. It is equivalent to adding a set of securities with the opposite characteristics from those included in the set when short sales are not allowed.

SECURITY SELECTION WITH A PURCHASABLE INDEX

Oftentimes the index used in the single-index model is a portfolio of securities. For example, the index could be the S&P index. If the portfolio used as an index is an asset in which the investor is considering investing (buy an index fund), then the simple rules described earlier are even simpler. As shown in Appendix E, in this case, Equation (9.3) collapses to

$$Z_i = \frac{\alpha_i'}{\sigma_{ei}^2}$$

where

$$\alpha_i' = \bar{R}_i - \left[R_F + \beta_i\left(\bar{R}_m - R_F\right)\right]$$

and the subscript m designates the index.

Once again, the amount to invest in any asset involves dividing each Z_i by the sum of the Z_is. The preceding expression, which works only if short sales are allowed, was first derived by Treynor and Black (1973). The intuition is that a mixture of a riskless asset and the index having the same beta as asset i would have an expected return of $R_F + \beta_i(\bar{R}_m - R_F)$. Thus, if asset i has a higher mean return than a passive mixture with the same beta $\alpha_i' > 0$, it should be held long. If it has a lower expected return than a passive mixture with the same beta $\alpha_i' < 0$, it should be sold short.[5]

[5]Because the riskless asset has a beta of zero and the index a beta of 1, the combination of a riskless asset and the portfolio with the same beta as asset i would involve investing β_i in the index portfolio and 1 minus β_i in the riskless portfolio. This has an expected return of $(1 - \beta_i)R_F + \beta_i\bar{R}_m$, or rearranging, $R_F + \beta_i(\bar{R}_m - R_F)$. This is the term in the brackets in the definition of α_i'.

Constructing an Efficient Frontier

The procedure just described assumes the existence of a riskless lending and borrowing rate. It produces the composition of the optimal portfolio that lies at the point where a ray passing through the riskless asset is tangent to the efficient frontier in expected return standard deviation space. If the investor does not wish to assume the existence of a riskless asset, then it is necessary to derive the full efficient frontier.

Two cases need to be analyzed: when short sales are allowed and when they are forbidden. If short sales are allowed, then, as was shown in Chapter 6, the full efficient frontier can be constructed from combinations of any two portfolios that lie on the efficient frontier. The composition of two portfolios on the efficient frontier can be found easily by assuming two different values for R_F and repeating the procedure just described for each. From these two efficient portfolios, the full frontier can be traced. The efficient frontier is a little more difficult to determine when short sales are not allowed.

The brute force solution is to solve the portfolio composition problem for a large number of values of R_F and, thus, approximate the full efficient frontier. An alternative procedure that solves directly for the R_F associated with each corner portfolio is described in Elton, Gruber, and Padberg (1978).[6] Because the frontier between corner portfolios can be found as combinations of corner portfolios, this procedure allows the full efficient frontier to be easily traced out.

THE CONSTANT CORRELATION MODEL

We now present and demonstrate the use of simple procedures for selecting optimum portfolios when the constant correlation model is accepted as the best way to forecast correlation coefficients. The reader will recall from earlier chapters that the constant correlation model assumes that the correlation between all pairs of securities is the same. The procedures assuming a constant correlation coefficient exactly parallel those presented for the case of the single-index model. Once again, the derivation of these procedures and the proof that they are, indeed, optimum is left for Appendix B at the end of this chapter.

If the constant correlation model is accepted as describing the comovement between securities, then all securities can be ranked by their excess return to standard deviation. To be precise, if σ_i is the standard deviation of the return on security i, then a security's desirability is determined by

$$\frac{\left(\bar{R}_i - R_F\right)}{\sigma_i}$$

Notice that we are still ranking on the basis of excess return to risk; but standard deviation has taken the place of beta as the relevant risk measure.[7] This ratio provides an ordering of securities for which the top-ranked securities are purchased and the lower-ranked securities are not held in the case of short sales prohibited or are sold short if such sales are allowed. Once again, there is a unique cutoff rate.

[6]Recall that a corner portfolio is one in which a security either enters the efficient set or is deleted from the efficient set as we move along the efficient frontier.

[7]In Chapter 18 we see that excess return to standard deviation, like excess return to beta, has been used as a technique for ranking portfolios.

Ranking and Selecting from among Securities—Short Sales Not Allowed

We illustrate the manner in which an optimal portfolio can be designed with a simple example presented in Table 9.6. First, as has been done in Table 9.6, all stocks are ranked by excess return to standard deviation. Then, the optimal value of C_i, called C^*, is calculated and all stocks with higher excess returns to standard deviation are included in the optimal portfolio. All stocks with lower excess returns to standard deviation are excluded. For the moment, accept that C^* equals 5.25. Shortly we will discuss how to calculate it. Because securities 1 through 3 have higher excess returns to standard deviations, they are included in the optimum portfolio. Securities 4 through 12 have excess returns to standard deviation below 5.25 and, hence, are not included in the optimal portfolio.

Setting the Cutoff Rate

The procedure for setting the cutoff rate is directly analogous to that presented for the case of the single-index model. First, we need a general expression for C_i, where i represents the fact that the first i securities are included in the computation of C_i. As shown in Appendix D at the end of this chapter, C_i can be found from

$$C_i = \frac{\rho}{1-\rho+i\rho}\sum_{j=1}^{i}\frac{\overline{R}_j - R_F}{\sigma_j}$$

where ρ is the correlation coefficient—assumed constant for all securities. The subscript i indicates that C_i is calculated using data on the first i securities.

Just as in the single-index model case, we have determined the appropriate level of the cutoff rate C^* when we have found a C_i such that

1. all stocks ranked 1 through i have a value of excess return to standard deviation lower than C_i
2. all stocks ranked $i + 1$ through N have a value of excess return to standard deviation lower than C_i

Table 9.6 Data to Determine Ranking $R_F = 5\%$

Security No. i	Expected Return $\overline{R}_i$	Excess Return $\overline{R}_i - R_F$	Standard Deviation σ_i	Excess Return to Standard Deviation $\frac{(\overline{R}_i - R_F)}{\sigma_i}$
1	29	24	3	8.0
2	19	14	2	7.0
3	29	24	4	6.0
4	35	30	6	5.0
5	14	9	2	4.5
6	21	16	4	4.0
7	26	21	6	3.5
8	14	9	3	3.0
9	15	10	5	2.0
10	9	4	2	2.0
11	11	6	4	1.5
12	8	3	3	1.0

Tables 9.6 and 9.7 present an example and some of the intermediate calculations needed to design an optimal portfolio. Examine the two columns at the extreme right of Table 9.7. Note that only for a value of $C_i = C_3$ do all stocks 1 to i have higher excess returns to standard deviation and all stocks $i + 1$ to 12 have lower excess return to standard deviation. Thus $C^* = C_3 = 5.25$.

As we show in Appendix B at the end of this chapter, the optimum amount to invest in any security is

$$X_i = \frac{Z}{\sum_{j=1}^{N} Z_j}$$

where

$$Z_i = \frac{1}{(1-\rho)\sigma_i}\left[\frac{\bar{R}_i - R_F}{\sigma_i} - C^*\right]$$

For our example we have

$$Z_1 = \frac{1}{1.5}\left[8 - \frac{21}{4}\right] = \frac{11}{6} = \frac{44}{24}$$

$$Z_2 = \frac{1}{1}\left[7 - \frac{21}{4}\right] = \frac{7}{4} = \frac{42}{24}$$

$$Z_3 = \frac{1}{2}\left[6 - \frac{21}{4}\right] = \frac{3}{8} = \frac{9}{24}$$

Table 9.7 Determining the Cutoff Rate $\rho = 0.5$

Security No. i	$\frac{\rho}{1-\rho+i\rho}$	$\sum_{j=1}^{i} \frac{R_j - R_F}{\sigma_j}$	C_i	$\frac{R_i - R_F}{\sigma_i}$
1	$\frac{1}{2}$	8	$\frac{8}{2} = 4$	8
2	$\frac{1}{3}$	15	$\frac{15}{3} = 5$	7
3	$\frac{1}{4}$	21	$\frac{21}{4} = 5.25$	6
4	$\frac{1}{5}$	26	$\frac{26}{5} = 5.2$	5
5	$\frac{1}{6}$	30.5	$\frac{30.5}{6} = 5.08$	4.5
6	$\frac{1}{7}$	34.5	$\frac{34.5}{7} = 4.93$	4
7	$\frac{1}{8}$	38	$\frac{38}{8} = 4.75$	3.5
8	$\frac{1}{9}$	41	$\frac{41}{9} = 4.56$	3
9	$\frac{1}{10}$	43	$\frac{43}{10} = 4.3$	2
10	$\frac{1}{11}$	45	$\frac{45}{11} = 4.09$	2
11	$\frac{1}{12}$	46.5	$\frac{46.5}{12} = 3.88$	1.5
12	$\frac{1}{13}$	47.5	$\frac{47.5}{13} = 3.65$	1

Dividing each Z_i by the sum of the Z_is gives the optimum amount to invest in each security. This calculation results in

$$X_1 = \frac{44}{44+42+9} = \frac{44}{95} \text{ or } 46.3\%$$
$$X_2 = \frac{42}{95} \text{ or } 44.2\%$$
$$X_3 = \frac{9}{95} \text{ or } 9.5\%$$

Short Sales Allowed

If short sales are allowed, then, as in the single-index case, all stocks will either be held long or sold short. This suggests, once again, that C^* should include all stocks, and this is correct. The C^* when all stocks are included is $C^* = C_{12} = 3.65$. Once again, C^* is the cutoff rate that separates securities that are purchased long from those that are sold short. In this example, $C^* = 3.65$ implies that the first six securities are purchased long and securities 7 to 12 are sold short. The optimum amount to invest in any security is given by the same formula Equation (9.4) with C^* defined to incorporate all securities.

OTHER RETURN STRUCTURES

We have presented two simple ranking devices based on different correlation structures. As discussed in the last two chapters, there are a number of other models for estimating the covariance structure. For each of these other structures a simple ranking device exists; the references listed at the end of the chapter show where. However, a few comments are in order. There are two types of models for estimating correlation structure: index models and group models. The single- and multi-index models are examples of the former, whereas constant correlation and multi-group models are examples of the latter.

For index models, the ranking is done by excess return to beta. This is true for both single- and multi-index models. However, the cutoff rate for multi-index models is different than the cutoff rate for single-index models. For example, assume a multi-index model where securities are related to a general market index and an industry index. In this model, the cutoff rate is different for each industry but depends on the members of all industries.

If a multi-group model is employed, then the ranking is always in terms of excess return to standard deviation. The cutoff rate varies from group to group and depends on which securities are included and in which groups.

Beta is important in index models because it is a measure of the securities' contribution to the risk of the portfolio. In multigroup or constant correlation models, the contribution to portfolio risk depends on the standard deviation, and hence standard deviation is the risk measure in the portfolios.

AN EXAMPLE

Let us return to the problem analyzed in Chapter 7. The problem involved an allocation among five common stock funds. The input data were

FUND	$\bar{R}_i$	β_i	$\sigma^2_{\epsilon i}$	$R_F = 5\%$
1. Small stock	23.5	1.4	65	

2.	Value	14	0.8	20
3.	Growth	20.75	1.3	45
4.	Large capitalization	12.05	0.9	24
5.	Special situation	13.95	1.1	45

Utilizing the simple rules discussed earlier, we can complete their rank in order of desirability:

FUND	$\frac{\overline{R}_i - R_F}{\beta_i}$
1.	13.21
2.	11.25
3.	12.12
4.	7.83
5.	8.14

Thus the ranking is 1, 3, 2, 5, 4. Calculating a cutoff rate assuming two securities in the optimum portfolio (1 + 3) yields

$$C^* = 11.82$$

This is optimum because securities 1 and 3 are above the cutoff and 2, 4, and 5 are below. Security 2 would be the next to enter. It is 0.57 below the cutoff. Thus, if the management is confident that the value fund has an expected return of 14 and the beta is estimated correctly, security 3 should not enter. Securities 4 and 5 are much further below the cutoff. Security 4 is 3.99 and 5 is 3.68 below. These are sufficiently far from the cutoff that reasonable adjustments in inputs are unlikely to lead to their inclusion. However, management might well wish to refine their estimates for securities 1, 2, and 3.

Computing the optimum proportions with no short sales, we have

$$Z_1 = \frac{1.4}{65}[13.21 - 11.82] = 0.02994$$

$$Z_3 = \frac{1.3}{45}[12.12 - 11.82] = 0.00867$$

and

$$X_1 = \frac{0.02994}{0.03861} = 0.775$$

$$X_2 = 0.225$$

It is left as an exercise for the reader to show that this solution is identical to the solution obtained using the technique discussed in Chapter 6.

CONCLUSION

In this chapter we have discussed several simple rules for optimal portfolio selection. These simple ranking devices allow the portfolio manager to quickly and easily determine the optimum portfolio. Furthermore, the manager uncertain about some of the estimates can easily manipulate them to determine if reasonable changes in the estimates lead to a different selection decision. The existence of a cutoff rate allows the manager to quickly determine if a new security should or should not be included in the portfolio.

Finally, the existence of simple ranking devices makes clear the characteristics of a security that are important and why a security is included, or excluded, from a portfolio.

APPENDIX A

SINGLE-INDEX MODEL—SHORT SALES ALLOWED

In this appendix we derive the simple ranking device when the investor is allowed to short sell securities and where he wishes to act as if the single-index model adequately reflects the correlation structure between securities. As we showed in Chapter 6, if the investor wishes to assume a riskless lending and borrowing rate, then he can obtain an optimum portfolio by solving a system of simultaneous equations. If, conversely, he desires to trace out the full efficient frontier, then he must solve this same system of simultaneous equations for two risk-free rates. This to determine the characteristics of any two efficient portfolios and to trace out the efficient frontier. The system of simultaneous equations the investor solves is

$$\bar{R}_i - R_F = Z_i\sigma_i^2 + \sum_{\substack{j=1 \\ j\neq i}}^{N} Z_j\sigma_{ij} \qquad i = 1,\ldots,N \tag{A.1}$$

where

$\bar{R}_i$ is the expected return of security i

R_F is the return on the riskless asset

σ_i^2 is the variance of security i

σ_{ij} is the covariance between securities i and j

Z_i is proportional to the amount invested in security i

From Chapter 7 we know that if the single-index model is used to describe the structure of security returns, then the covariance between securities i and j is $\beta_i\beta_j\sigma_m^2$ and the variance of security i is $\beta_i^2\sigma_m^2 + \sigma_{ei}^2$. Substituting these relationships that hold for the single-index model into the general system of simultaneous equations, (A.1), yields

$$\bar{R}_i - R_F = Z_i\left(\beta_i^2\sigma_m^2 + \sigma_{ei}^2\right) + \sum_{\substack{j=1 \\ j\neq i}}^{N} Z_j\beta_i\beta_j\sigma_m^2 \qquad i = 1,\ldots,N$$

Look at the summation term. If $j = i$, it would be $Z_i\beta_i\beta_j\sigma_m^2$. But this is exactly the first term on the right-hand side of the equality sign. Eliminating the $j \neq i$ underneath the summation sign by incorporating the term $Z_i\beta_i\beta_i\sigma_m^2$ within it yields

$$\bar{R}_i - R_F = Z_i\sigma_{ei}^2 + \sum_{j=1}^{N} Z_j\beta_i\beta_j\sigma_m^2 \qquad i = 1,\ldots,N$$

Solving for Z_i and taking the constants outside the summation yields

$$Z_i = \frac{\bar{R}_i - R_F}{\sigma_{ei}^2} - \frac{\beta_i\sigma_m^2}{\sigma_{ei}^2}\sum_{j=1}^{N} Z_j\beta_j \qquad i = 1,\ldots,N \tag{A.2}$$

This can be written as

$$Z_i = \frac{\beta_i}{\sigma_{ei}^2}\left[\frac{\overline{R}_i - R_F}{\beta_i} - C^*\right] \qquad i = 1,\ldots,N$$

where

$$C^* = \sigma_m^2 \sum_{j=1}^{N} Z_j \beta_j \tag{A.3}$$

This is the equation presented in the text. To get the C^* in terms of known variables, we must express (A.2) and (A.3) in terms that do not invoke

$$\sum_{j=1}^{N} Z_j \beta_j$$

To do so, first multiply Equation (A.2) by β_i and sum over all values of $i = 1, \ldots, N$. This yields

$$\sum_{j=1}^{N} Z_j \beta_j = \sum_{j=1}^{N} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2} - \sigma_m^2 \sum_{j=1}^{N} \frac{\beta_j^2}{\sigma_{ej}^2} \sum_{j=1}^{N} Z_j \beta_j$$

Notice that the term

$$\sum_{j=1}^{N} Z_j \beta_j$$

is found on both the left-hand and right-hand sides of the equation. Solving for this yields

$$\sum_{j=1}^{N} Z_j \beta_j = \frac{\displaystyle\sum_{j=1}^{N} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \displaystyle\sum_{j=1}^{N} \frac{\beta_j^2}{\sigma_{ej}^2}}$$

From Equation (A.3) we see that

$$C^* = \frac{\sigma_m^2 \displaystyle\sum_{j=1}^{N} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \displaystyle\sum_{j=1}^{N} \frac{\beta_j^2}{\sigma_{ej}^2}}$$

The alternative form for C_i [Equation (9.2)] employed in the text can be derived from Equation (A.3). From Equation (A.3) we see that

$$C^* = \sigma_m^2 \sum_{j=1}^{N} Z_j \beta_j$$

We also note from Chapter 6 that Z_j is proportional to the optimal fraction of the portfolio the investor should hold in each stock X_j. The constant is equal to the ratio of the excess return of the optimal portfolio to the variance of its return. Thus

$$C^* = \sigma_m^2 \sum_{j=1}^{N} \frac{\bar{R}_P - R_F}{\sigma_P^2} X_j \beta_j$$

Recognizing

$$\sum_{j=1}^{N} X_j \beta_j$$

as the beta on the investor's portfolio

$$C^* = \left(\bar{R}_P - R_F\right) \beta_P \frac{\sigma_m^2}{\sigma_P^2}$$

Dividing and multiplying the equation by β_i and recognizing that $\beta_i \beta_p \sigma_m^2$ is cov(ip) under the assumption of the single-index model, we have

$$C^* = \left(\bar{R}_P - R_F\right) \frac{\text{cov}(ip)}{\sigma_P^2} \frac{1}{\beta_i} = \frac{\beta_{ip}}{\beta_i} \left(\bar{R}_P - R_F\right)$$

where β_{ip} is the regression coefficient of the return on security i to the return on portfolio p.

APPENDIX B

CONSTANT CORRELATION COEFFICIENT—SHORT SALES ALLOWED

In this appendix we derive the simple ranking devices discussed in the text when the investor believes that the constant correlation coefficient adequately describes the structure of security returns. Once again, we utilize the result shown in Chapter 6 that the efficient frontier can be determined by solving a system of simultaneous equations. The system of simultaneous equations is

$$\bar{R}_i - R_F = Z_i \sigma_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{N} Z_i \sigma_{ij} \qquad i = 1, \ldots, N \tag{B.1}$$

If the constant correlation model holds, then $\sigma_{ij} = \rho \sigma_i \sigma_j$. Note that the correlation coefficient between stocks i and j is by assumption the same for all i and j. Making the substitution into (B.1) yields

$$\bar{R}_i - R_F = Z_i \sigma_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{N} Z_j \rho \sigma_i \sigma_j \qquad i = 1, \ldots, N$$

If $j = i$, then the term in the summation is $Z_i \rho \sigma_i \sigma_i$. Adding this to the summation and subtracting the same term yields

$$\bar{R}_i - R_F = Z_i \sigma_i^2 - Z_i \rho \sigma_i \sigma_i + \sum_{j=1}^{N} Z_j \rho \sigma_i \sigma_j \qquad i = 1, \ldots, N$$

Solving for Z_i yields

$$Z_i(1-\rho)\sigma_i^2 = \bar{R}_i - R_F - \rho\sigma_i \sum_{j=1}^{N} Z_j\sigma_j \qquad i = 1,\ldots,N$$

or

$$Z_i = \frac{1}{(1-\rho)\sigma_i}\left[\frac{\bar{R}_i - R_F}{\sigma_i} - C^*\right] \qquad i = 1,\ldots,N \tag{B.2}$$

where

$$C^* = \rho\sum_{j=1}^{N} Z_j\sigma_j$$

This is the equation used in the text. To express C^* in known terms, multiply (B.2) by σ_i and ρ and add up the N equations. This yields

$$C^* = \rho\sum_{j=1}^{N} Z_j\sigma_j = \frac{\rho}{1-\rho}\sum_{j=1}^{N}\frac{\bar{R}_j - R_F}{\sigma_j} - \frac{N\rho C^*}{1-\rho}$$

Solving for C^*,

$$C^*\left(1 + \frac{N\rho}{1-\rho}\right) = \frac{\rho}{1-\rho}\sum_{j=1}^{N}\frac{\bar{R}_j - R_F}{\sigma_j}$$

or

$$C^* = \left(\frac{\rho}{1-\rho}\right)\left(\frac{1-\rho}{1-\rho+N\rho}\right)\sum_{j=1}^{N}\frac{\bar{R}_j - R_F}{\sigma_j} = \frac{\rho}{1-\rho+N\rho}\sum_{j=1}^{N}\frac{\bar{R}_j - R_F}{\sigma_j}$$

APPENDIX C

SINGLE-INDEX MODEL—SHORT SALES NOT ALLOWED

In this appendix we derive simple ranking rules when the investor wishes to act as if the single-index model is a reasonable method of describing the structure of security returns. In Chapter 6 we showed that if we could find a solution that met the Kuhn–Tucker conditions, then we could be certain we had the optimum portfolio. In this appendix we show that our simple ranking procedure does, in fact, lead to a solution that meets the Kuhn–Tucker conditions.

The Kuhn–Tucker conditions were

1. $\bar{R}_i - R_F = Z_i\sigma_i^2 + \sum_{\substack{j=1\\ j\neq i}}^{N} Z_j\sigma_{ij} - M_i \qquad i = 1,\ldots,N.$

2. $Z_iM_i = 0 \qquad i = 1, ..., N.$

3. $Z_i \geq 0$ and $M_i \geq 0 \qquad i = 1, ..., N.$ (C.1)

where M_i is a variable added to make Equation (C.1) an equality.

If the single-index model is assumed to adequately describe the return structure, then

$$\sigma_{ij} = \beta_i \beta_j \sigma_m^2 \quad \text{and} \quad \sigma_i^2 = \beta_i^2 \sigma_m^2 + \sigma_{ei}^2$$

Substituting this into the first Kuhn–Tucker condition yields

$$\overline{R}_i - R_F = Z_i\left(\beta_i^2 \sigma_m^2 + \sigma_{ei}^2\right) + \sum_{\substack{j=1 \\ j \neq i}}^{N} Z_j \beta_i \beta_j \sigma_m^2 - M_i \qquad i = 1, \ldots, N$$

Once again, noting that when $j = i$, the term in the summation would be $Z_i\beta_i\beta_i\sigma_m^2$, and this is the first term on the right-hand side of the equality. Incorporating this term into the summation, we have

$$\overline{R}_i - R_F = Z_i \sigma_{ei}^2 + \beta_i \sigma_m^2 \sum_{j=1}^{N} Z_j \beta_j - M_i \qquad i = 1, \ldots, N$$

If the security is not in the optimum portfolio, then $Z_j = 0$. Thus the summation only has to include the Z_i and β_i for those securities in the optimum portfolio. We will call the set of securities in the optimum set *k*. Furthermore, we will use the symbol

$$\sum_{j \epsilon k}$$

to indicate that the summation is to include all securities in the optimum. Rewriting the equation yields

$$\overline{R}_i - R_F = Z_i \sigma_{ei}^2 + \beta_i \sigma_m^2 \sum_{j \epsilon k} Z_j \beta_j - M_i \qquad i = 1, \ldots, N \tag{C.2}$$

Examine conditions 2 and 3. Condition 3 says that Z_i and M_i must each be either zero or positive. Condition 2 states that their product must be zero. Thus, if Z_i is positive, M_i must be zero. For any security included in the optimum, Z_i is positive. Hence, we can drop the M_i for included securities (those in set *k*). Setting $M_i = 0$ in Equation (C.2) yields

$$\overline{R}_i - R_F = Z_i \sigma_{ei}^2 + \beta_i \sigma_m^2 \sum_{j \epsilon k} Z_j \beta_j \qquad \text{for } i \epsilon k$$

or

$$Z_i = \frac{\beta_i}{\sigma_{ei}^2}\left[\frac{\overline{R}_i - R_F}{\beta_i} - \sigma_m^2 \sum_{j \epsilon k} Z_j \beta_j\right] \qquad \text{for } i \epsilon k \tag{C.3}$$

We can eliminate

$$\sum_{j \epsilon k} Z_j \beta_j$$

by multiplying (C.3) by β_j and summing over set *k*:

$$\sum_{j \epsilon k} Z_j \beta_j = \sum_{j \epsilon k} \frac{\left(\overline{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2} - \sigma_m^2 \sum_{j \epsilon k} \frac{\beta_j^2}{\sigma_{ej}^2} \sum_{j \epsilon k} Z_j \beta_j$$

Rearranging

$$\sum_{j\epsilon k} Z_j \beta_j = \frac{\sum_{j\epsilon k} \frac{\left(\bar{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \sum_{j\epsilon k} \frac{\beta_j^2}{\sigma_{ej}^2}}$$

(C.3) can be written as

$$Z_i = \frac{\beta_i}{\sigma_{ei}^2}\left[\frac{\bar{R}_i - R_F}{\beta_i} - C^*\right] \qquad i\epsilon k \tag{C.4}$$

where

$$C^* = \sigma_m^2 \sum_{j\epsilon k} Z_j \beta_j = \frac{\sigma_m^2 \sum_{j\epsilon k} \frac{\left(\bar{R}_j - R_F\right)\beta_j}{\sigma_{ej}^2}}{1 + \sigma_m^2 \sum_{j\epsilon k} \frac{\beta_j^2}{\sigma_{ej}^2}}$$

This is the expression utilized in the text.

Let us see how to determine a portfolio that meets the Kuhn–Tucker conditions. First, condition 2 ($Z_i M_i = 0$) is met by construction. M_i was set to zero for all securities included in the optimum portfolio, those with $Z_i > 0$. For those not included in the optimum, $Z_i = 0$, guaranteeing $Z_i M_i = 0$.

Now consider the first and third conditions. Assume we have found a set of securities for which Z_i as determined by (C.4) is greater than zero for securities in the set and less than zero for securities not in the set.

For securities in the set, Equation (C.4) is equivalent to condition 1 if $M_i = 0$. $Z_i > 0$ and $M_i = 0$ meet condition 3. Thus conditions 1 and 3 are met.

For securities not in this set, (C.4) is not equivalent to condition 1. However, comparing these two shows that $M_i > 0$ will make condition 1 hold, and also Z_i is equal to zero so that condition 3 holds.

Thus the Kuhn–Tucker conditions will be met if a set k can be determined for which (C.4) is positive for members of the set and negative for securities not in the set.

Examine (C.4). C^* is a constant. Assume for the moment that $\beta_i > 0$. Then the term outside the brackets is positive. The term in the brackets is positive if $(\bar{R}_i - R_F)/\beta_i > C^*$ and is negative if $(\bar{R}_i - R_F)/\beta_i < C^*$. The procedure discussed in the text assures that this will occur.

APPENDIX D

CONSTANT CORRELATION COEFFICIENT—SHORT SALES NOT ALLOWED

The analysis in this section closely parallels the analysis of the last section. Once again, if the Kuhn–Tucker conditions are met, then the solution is an optimum. The Kuhn–Tucker conditions are as shown in (C.1). If an investor wishes to act as if the return structure is adequately described by the assumption of a constant correlation coefficient, then the

covariance terms are $\sigma_{ij} = \rho\sigma_i\sigma_j$. Making this substitution into the first Kuhn–Tucker condition and adding and subtracting $\rho\sigma_i\sigma_i Z_i$ to eliminate $j \neq i$ under the summation sign yields

$$\textbf{1.} \quad \bar{R}_i - R_F = Z_i\sigma_i^2(1-\rho) + \sum_{j=1}^{N} Z_j\rho\sigma_i\sigma_j - M_i \qquad i = 1,\ldots,N$$

$$\textbf{2.} \quad Z_iM_i = 0, \qquad i = 1, ..., N$$

$$\textbf{3.} \quad Z_i \geq 0 \text{ and } M_i \geq 0 \qquad i = 1, ..., N \tag{D.1}$$

The same considerations hold here as did in Appendix C. If set k is the set of included securities, then $Z_i = 0$ for securities not in set k and thus

$$\sum_{j\epsilon k} Z_j\sigma_j = \sum_{j=1}^{N} Z_j\sigma_j$$

Furthermore, if $Z_i > 0$, then $M_i = 0$ so that $M_i = 0$ for set k. Using these two observations, Equation (D.1) becomes

$$\bar{R}_i - R_F = Z_i\sigma_i^2(1-\rho) + \rho\sigma_i \sum_{j\epsilon k} Z_j\sigma_j \qquad i\epsilon k \tag{D.2}$$

Rearranging and solving for Z_i,

$$Z_i = \frac{1}{(1-\rho)\sigma_i}\left[\frac{\bar{R}_i - R_F}{\sigma_i} - \rho\sum_{j\epsilon k} Z_j\sigma_j\right] \tag{D.3}$$

We can eliminate

$$\sum_{j\epsilon k} Z_j\sigma_j$$

by multiplying each equation by σ_i and then adding together all the equations in set k. This yields

$$\sum_{j\epsilon k} Z_j\sigma_j = \frac{1}{1-\rho}\left[\sum_{j\epsilon k}\frac{\bar{R}_j - R_F}{\sigma_j} - \rho N_k\sum_{j\epsilon k} Z_j\sigma_j\right]$$

where N_k is the number of securities in k. Rearranging,

$$\sum_{j\epsilon k} Z_j\sigma_j = \frac{1}{1-\rho}\left(\frac{1-\rho}{1-\rho+\rho N_k}\right)\sum_{j\epsilon k}\frac{\bar{R}_j - R_F}{\sigma_j}$$

Thus (D.3) becomes

$$Z_i = \frac{1}{(1-\rho)\sigma_i}\left[\frac{\bar{R}_i - R_F}{\sigma_i} - \phi_k\right] \qquad i\epsilon k$$

$$\phi_k = \rho\sum_{j\epsilon k} Z_j\sigma_j = \frac{\rho}{1-\rho+\rho N_k}\sum_{j\epsilon k}\frac{\bar{R}_j - R_F}{\sigma_j} \tag{D.4}$$

The same considerations hold here as did in Appendix C. Namely, if (D.4) is positive for members of set k and negative for all other securities, the Kuhn–Tucker conditions are met. The procedures discussed in the text lead to this solution.

APPENDIX E

SINGLE-INDEX MODEL, SHORT SALES ALLOWED, AND A MARKET ASSET

If one can buy a portfolio that exactly replicates the index used in the single-index model, the solution is simpler. In fact investors can often replicate the index. For example, the Standard and Poor's (S&P) index is often used as the index in the single-index model, and an investor can buy an index fund matching the S&P index.

We will now examine this case. Let the subscript m represent this asset. Furthermore, note that portfolio m regressed on itself has zero residual risk and a slope of 1. Thus $\sigma^2_{em} = 0$ and $\beta_m = 1$. With these substitutions the equation above (A.2) becomes $\bar{R}_m - R_F = \sigma^2_m \sum_{j=1}^{N} \beta_j Z_j$, and thus the cutoff rate in (A.3) is $C^* = \bar{R}_m - R_F$. Substituting this into (9.3) results in

$$Z_i = \frac{\beta_i}{\sigma^2_{ei}}\left(\frac{\bar{R}_i - R_F}{\beta_i} - \left(\bar{R}_m - R_F\right)\right)$$

or

$$Z_i = \frac{1}{\sigma^2_{ei}}\left[\bar{R}_i - R_F - \beta_i\left(\bar{R}_m - R_F\right)\right]$$

Defining α'_i as the term in the brackets, we have $Z_i = \frac{\alpha'_i}{\sigma^2_{ei}}$, which is the expression shown in the text. This expression does not hold when short sales are not allowed. In particular, the solution when short sales are not allowed does not involve holding long all securities with a positive Z_i.

QUESTIONS AND PROBLEMS

1. Given the following data: $\sigma^2_m = 10$

Security Number	Expected Return	Beta	σ^2_{ei}
1	15	1.0	30
2	12	1.5	20
3	11	2.0	40
4	8	0.8	10
5	9	1.0	20
6	14	1.5	10

What is the optimum portfolio assuming no short sales if $R_F = 5\%$?

2. What is the optimum portfolio assuming short sales if $R_F = 5\%$ and the data from Problem 1 are used?

3. Using the data from Problem 1, what is the optimum portfolio assuming short sales are allowed but riskless lending and borrowing are forbidden?

4. Given the following data

Security Number	Expected Return	Standard Deviation
1	15	10
2	20	15
3	18	20
4	12	10
5	10	5
6	14	10
7	16	20

What is the optimum portfolio assuming no short sales if $R_F = 5\%$ and $\rho = 0.5$?

5. What is the optimum portfolio assuming short sales if $R_F = 5\%$ and $\rho = 0.5$? Use the data in Problem 4.

6. What is the optimum portfolio assuming short sales but no riskless lending and borrowing with $\rho = 0.5$ for all pairs of securities? Use the data in Problem 4.

BIBLIOGRAPHY

1. Alexander, Gordon J., and Resnick, Bruce G. "More on Estimation Risk and Simple Rules for Optimal Portfolio Selection," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 125–134.
2. Bawa, Vijay, Elton, Edwin J., and Gruber, Martin J. "Simple Rules for Optimal Portfolio Selection in a Stable Paretian Market," *Journal of Finance*, **34,** No. 2 (June 1979), pp. 1041–1047.
3. Chen, Son-Nan, and Brown, Stephen J. "Estimation Risk and Simple Rules for Optimal Portfolio Selection," *Journal of Finance*, **38,** No. 4 (Sept. 1983), pp. 1087–1094.
4. Elton, Edwin J., Gruber, Martin J., and Padberg, Manfred W. "Simple Criteria for Optimal Portfolio Selection," *Journal of Finance*, **XI,** No. 5 (Dec. 1976), pp. 1341–1357.
5. ——. "Simple Rules for Optimal Portfolio Selection: The Multi Group Case," *Journal of Financial and Quantitative Analysis*, **XII,** No. 3 (Sept. 1977), pp. 329–345.
6. ——. "Simple Criteria for Optimal Portfolio Selection: Tracing Out the Efficient Frontier," *Journal of Finance*, **XIII,** No. 1 (March 1978), pp. 296–302.
7. ——. "Optimal Portfolios from Simple Ranking Devices," *Journal of Portfolio Management*, **4,** No. 3 (Spring 1978), pp. 15–19.
8. ——. "Simple Criteria for Optimal Portfolio Selection with Upper Bonds," *Operations Research*, **8** (Nov.–Dec. 1978), pp. 952–967.
9. ——. "Simple Criteria for Optimal Portfolio Selection: The Multi-Index Case," in Edwin J. Elton and Martin J. Gruber (eds.), *Portfolio Theory: 25 Years Later* (Amsterdam: North-Holland, 1979).
10. Frankfurter, George M., and Lamourex, Christopher G. "The Relevance of the Distributional Form of Common Stock Returns to the Construction of Optimal Portfolios," *Journal of Financial and Quantitative Analysis*, **22,** No. 4 (Dec. 1987), pp. 505–511.
11. Green, Richard C., and Hollifield, Burton. "When Will Mean–Variance Efficient Portfolios Be Well Diversified?" *Journal of Finance*, **47,** No. 5 (Dec. 1992), pp. 1785–1809.
12. Kwan, Clarence C. Y. "Portfolio Analysis Using Single Index, Multi-Index, and Constant Correlation Models: A Unified Treatment," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1469–1484.
13. Lee, Sang, and Lerro, A. J. "Optimizing the Portfolio Selection for Mutual Funds," *Journal of Finance*, **VIII,** No. 5 (Dec. 1973), pp. 1087–1101.
14. Mao, C. T. James. "Essentials of Portfolio Diversification Strategy," *Journal of Finance*, **V,** No. 5 (Dec. 1970), pp. 1109–1121.

15. Porter, Burr, and Bey, Roger. "An Evaluation of the Empirical Significance of Optimal Seeking Algorithms in Portfolio Selection," *Journal of Finance*, **IX,** No. 5 (Dec. 1974), pp. 1479–1490.
16. Sharpe, W. F. "Simple Strategies for Portfolio Diversification: Comment," *Journal of Finance*, **VII,** No. 1 (March 1972), pp. 127–129.
17. ——. "Simple Strategies for Portfolio Diversification: Comment, a Correction," *Journal of Finance*, **VII,** No. 3 (June 1972), p. 733.
18. Sharpe, William, and Stone, Bernell. "A Linear Programming Formulation of the General Portfolio Selection Model," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 4 (Sept. 1973), pp. 621–636.
19. Treynor, J., and Black, F. "Using Security Analysis to Improve Portfolio Selection," *Journal of Business*, **46,** No. 1 (1973), pp. 66–86.

Section 3

Selecting the Optimum Portfolio

10

Estimating Expected Returns

As discussed in earlier chapters, to implement modern portfolio theory, one must have estimates of future expected returns, variances, and covariances. Of these, the hardest to forecast is future expected returns. The valuation of a company's stock, as well as the valuation of the stock market as a whole, depends on the aggregate of all participants' expectations. Future returns are heavily dependent on how these expectations change over time. Because these expectations are unobservable, and there are so many diverse opinions, it is difficult to forecast the change in aggregate expectations. There is no magic formula for forecasting future expected returns. However, there are some general techniques that have proved helpful in the past. These are discussed in this chapter.

The chapter is divided into three sections. In the first section, we discuss forecasting return for asset categories like stocks or bonds. These forecasts are useful in aggregate asset allocation. In the second section, we discuss forecasting mean return for individual securities. In the final section, we discuss dealing with forecasts when the forecasts are discrete (e.g., buy, hold, sell) rather than continuous (e.g., expected return is 13.2%).

AGGREGATE ASSET ALLOCATION

Aggregate asset allocation deals with how much to invest in broad categories of securities. For example, in what proportions should an investor divide his assets among large capitalization stocks, international stocks, government bonds, corporate bonds, and real estate? This asset allocation problem is faced by almost all investors.

All pension plans must make asset allocation decisions. If a company is managing the pension plan for its employees, there will be a plan administrator who will typically employ outside portfolio managers to select individual securities. However, the plan administrator will decide how to split the plan assets among portfolio managers and asset types. Alternatively, if the participant is managing his or her own pension plan, then the participant has to decide how much to put in each asset category. Both of these choices are aggregate asset allocation decisions.

Aggregate asset allocation decisions are made by many other types of investment entities. Endowment managers, for example—the managers of the assets held by the

Metropolitan Museum of Art, the Kidney Foundation, or New York University—all have to decide how much to put in each asset category, whether they are investing directly or, more commonly, use outside managers to manage individual asset categories. Company savings or profit-sharing plans are also often managed in a similar fashion, where outside portfolio managers are used for securities selection and the plan manager's task is to allocate among these portfolio managers.

How can this decision be made? Expected returns for asset categories are usually estimated in a three-step procedure by determining

1. the normal return for the asset category
2. how much you expect returns in the next period to deviate from normal
3. the expected deviation of the particular manager hired to manage an asset category from the average for that category

The first two are discussed in this chapter. The third is the subject of Chapter 25 on performance evaluation. Before discussing procedures for allocating across asset categories, it is useful to discuss market timing.

Market Timing or Dynamic Asset Allocation

Numerous studies show how much money could be made if one bought stocks or bonds before these asset categories had large positive returns and sold them before periods when returns were negative. This strategy was traditionally called *market timing*; more recently, it has been referred to as *dynamic asset allocation.* The price of securities depends on the average beliefs of investors (where each dollar invested gets one vote). As an example, consider bonds. Bond prices depend on expectations about future interest rates, and consensus beliefs about future interest rates are impounded in today's bond prices. To successfully time bond returns, the manager has to not only forecast future interest rates different from the consensus but also be more accurate than the consensus in these forecasts. The researchers who have studied managers that market time have found little evidence that would suggest that managers can successfully market time.[1]

Market timing has a second difficulty. When a manager selects 100 securities for a portfolio, each selection is influenced by forecasts of the individual security's return. If there is some information in these forecasts, but each forecast has a large error, then the information may lead to a superior portfolio, even with large errors, since the manager's portfolio is an average across hundreds of forecasts and the errors in each forecast that are not systematic will tend to cancel out. Thus a manager who has access to superior forecasts of security returns should have better performance in most periods, and this should be detectable by an outside observer.

A market timing decision is a single forecast. If the manager has some ability to forecast future market movements, but with a large error (e.g., correct 53% of the time), the chances of being incorrect at any one point in time or several time periods in a row are high. It will take many years for random errors to cancel out and for an observer to have a reasonable chance of determining whether a manager has superior timing ability. In addition, the manager, to successfully market time, must take large positions in a subset of individual asset categories. This means that the portfolio will be less diversified and the risks of a large loss from a timing decision are greater than the risks resulting from selection. Pension plan administrators and endowment managers are normally unwilling

[1]See Comer (2006), Henriksson and Merton (1981), and Treynor and Muzay (1966).

to wait years to be able to determine if they have employed portfolio managers with superior timing skills. Of course, if they already employ a market timer, it will take years to determine that the timer is not doing a good job. In addition, if the plan administrators or endowment managers are engaged in timing themselves, their bosses are unlikely to be willing to wait years to be reasonably sure of an appropriate evaluation of their skills.

Therefore, because the weight of the evidence does not support an ability of managers to successfully time, because market timing portfolios have great risk, and due to an unwillingness to wait years to determine if a manager has superior ability, most administrators set target weights at fixed levels for alternative asset classes and allow small derivatives from the target weight over time.

Estimating Expected Returns

As indicated earlier, the first step toward determining the expected return to use as an input to the asset allocation process is to determine the normal returns of the asset. One approach is to simply use the long-term historical performance of the asset class as an input. This historical performance data at the asset-class level were first available from Ibbotson and Sinquefield (1976) and subsequently from Ibbotson Associates. These data make it possible to calculate the returns of U.S. equities from 1926 to the present. When the expected return of the stock market is constant through time, the longer the historical data series, the more precise is the estimate of the mean of the series. However, this stationarity in the expected returns to stocks cannot simply be assumed—particularly over long periods of war and peace, technological change, and varying macroeconomic conditons that might affect the performance of the stock market. Table 10.1 reports the performance of various U.S. investment classes. The difference between the returns on large-company stocks and the U.S. 30-day T-bill return is called the *equity premium*. It is the amount of return that investors demand for holding a risky security such as stocks, as opposed to a riskless security such as T-bills. The annual equity premium is about 8.29% over the 1926–2011 period. Because assets are priced based on their relative risk, it is generally believed that if one can estimate one asset category with relative accuracy, pricing other assets relative to that category is a superior way to forecast expected returns. The expected one-year return on a one-year Treasury bill is its yield to maturity—a quantity that can be observed in the market. Thus they are a useful benchmark for building expected returns for other categories. The equity premium is added to the current Treasury bill rate to form a forward-looking expected return for U.S. equities. This same principle can be applied to other asset classes as well. In equilibrium, risk assets must pay a premium over the riskless asset to induce investors to hold them.

Table 10.1

	Spread to 2011	
Return over Treasury Bills	1926–2011	1872–2011
Large company stocks	8.2	11.5
Small company stocks	12.9	15.4
Long-term corporate	2.8	9.3
Long-term government	2.3	9.6

History and the Equity Risk Premium

To estimate the equity risk premium over the longest possible time period, Goetzmann, Ibbotson, and Peng (2001) gathered stock performance data extending back to 1815 from the New York Stock Exchange (NYSE).[2] These results, presented in Table 10.2, indicate that the average equity risk premium from 1815 to 1925 as measured by the spread over U.S. government bonds was about 3.8%. This is significantly lower than the premium over the period following 1925, based on results presented in Table 10.3. The difference may be that U.S. government securities were not riskless over this early time period. In its early years, the United States was not the reliable borrower that it is today. The spread of stock returns over inflation was substantial in both time periods: 7% to 9% annually. This suggests that long-run forecasts of stock returns over inflation are reliably positive over nearly two centuries.

Dimson, Marsh, and Staunton (2002) have examined the returns to a number of countries from 1900 onward and found that a positive equity premium—whether measured net of government bonds, bills, or inflation—is the rule internationally rather than the exception. Over these long time periods, there have been extended periods of market decline as well as market growth. Crashes were not infrequent in U.S. capital market history. Sudden market declines occurred in 1837, 1907, 1929, 1971, 1987, 2000, and 2008. Despite these market dynamics, equities have provided a significant long-term positive premium.

Bayesian Models of Expected Returns

The estimation of expected returns from data, regardless of the length of the time series, always has the problem that the mean is estimated with statistical error.[3] Some researchers have addressed this issue by Bayesian methods. A point of departure in a Bayesian approach to portfolio choice considers that the distribution of return next period (the 'predictive distribution') includes uncertainty not only about the possible deviation of returns from expected values but also about these expected values themselves. As Klein and Bawa (1976) show, the fact that expected return is not known effectively adds to the risk faced by investors and leads them to choose portfolios that are more conservative (smaller investment in risky assets) than would be the case if they were to ignore uncertainty about values of expected returns. This additional risk is referred to as *estimation risk*.

The Bayesian approach uses reasonable priors about expected returns as a starting point for estimating expected returns from historical data. In Chapter 7 we noted that Bayesian techniques for estimating betas have proven useful in reducing out-of-sample error. In its most basic form, Bayesian estimation begins with a prior about the value to be estimated, in this case, the mean return of an asset class. This prior is updated by empirical data, and the posterior value, used in the mean variance analysis, is a mixture between the prior and the mean of the empirical data. This process "shrinks" the estimated mean toward the prior. A commonsense prior such as the assumption that stocks provide a higher risk premium than bonds is one example.

This concept was applied to the estimation of inputs to the asset allocation process by Brown (1976), who proposed Bayesian methods to address estimation risk.[4] Jorion (1986) used a related technique termed a James–Stein shrinkage estimator, which provides biased but greatly improved posteriors. The research on methodological improvements to the input estimation

[2]Evidence on historical risk premia is discussed in Goetzmann and Ibbotson (2006).
[3]Jorion (1992).
[4]Brown (1976) noted that this approach was equivalent up to a scalar transformation to a technique developed in Elton, Gruber, and Padberg. (1976).

process is ongoing. In recent work, Kan and Zhou (2007) further extend the Bayesian model and show significant progress in estimating out-of-sample optimal portfolios.

One interesting baseline for input estimation is one that assumes nothing at all is known about the risk, return, and covariances of the asset classes. With no statistical data to update beliefs about expectations, the optimum portfolio is a portfolio that allocates equally across all assets. For a portfolio with N assets classes, this corresponds to an equal-weighted portfolio with weights given by $1/N$. The simple logic is that, if you know nothing about any investment, naive diversification reduces risk. Brown (1976) and DeMiguel, Garlappi, and Uppal (2009) show that the $1/N$ portfolio of equities performs surprisingly well out of sample, particularly for small sample sizes, doing better than many other approaches, including reliance on historical inputs, to predict the ex post optimal portfolio. Although these results may not be applicable to allocation across multiple asset classes, as opposed to identifying an optimal stock-only portfolio, they nevertheless challenge the efficacy of standard approaches to input estimation. Kan and Zhou (2011) are more optimistic about statistical methods to selecting inputs. They find that a combination of the $1/N$ rule together with additional Bayesian methods performs even better. The broad lesson from this ongoing research is that, particularly in the case when there are many assets with similar expected returns, shrinkage toward a diffuse prior, or adjusting allocations away from extreme weights on few assets, has advantages.

Black and Litterman (1992) take a different approach to using Bayesian priors to improve input estimation. They use economic priors based on economic equilibrium arguments discussed in Chapter 13. They use a particular equilibrium model, the capital asset pricing model (CAPM). One implication of this model is that everyone should hold the market or world wealth portfolio. The CAPM makes strong assumptions about the composition of the global wealth portfolio. Black and Litterman argue that, if the CAPM is indeed a reasonable description of the world, then it makes a good prior. They ask, what set of inputs for the major asset classes will result in a portfolio that matches the weights of the world wealth portfolio? In their model, inputs that give a tangency portfolio dramatically different from that predicted by the CAPM are highly unlikely. This prior is then updated with empirical data. The Black–Litterman approach thus has the benefit of using financial theory to provide a way to reduce estimation risk.

Time Variation in Expected Returns

Is it possible to forecast periods for which a given asset class may deviate from the norm? Considerable research has been devoted to the question of whether it is possible to forecast stock market returns. Longer-term, multiple-year forecasts are most appropriate for the purposes of selecting inputs to the asset allocation process. There is some evidence that stock returns follow a mean-reverting process over multiple-year horizons[5] and also evidence that valuation ratios such as the earnings-price ratio and the dividend-price ratio may forecast deviations in the equity risk premium.[6] The dividend yield is a particularly compelling instrument for forecasting because, under the assumption of no uncertainty about interest rates, the current level of the stock market is equal to the discounted stream of future dividends it provides. When this future stream is perpetual and fixed, the discount rate for the market (its expected return) is equal to the ratio of the dividend over the price. Thus it would be logical to use this ratio as a predictor of future returns: when stocks have high prices compared to their current dividends, the model predicts a low future equity premium

[5]See Fama and French (1988) and Poterba and Summers (1988) for evidence on mean reversion in equity prices using various statistical measures.

[6]See Rozeff (1984), Campbell and Shiller (1988), and Goetzman and Jorion (1993, 1995).

Table 10.2

Summary Statistics for New York Stock Exchange Returns, U.S. Bond Yields, Call Money Rates, and Inflation 1792–1925

	Arithmetic Return	Geometric Return	Standard Deviation
Stocks	7.93%	6.99%	14.64%
Capital appreciation	1.91%		
Income	6.01%		
Bonds	4.17%	4.16%	4.17%
Commercial paper	7.62%	7.57%	3.22%
Inflation	0.85%	0.61%	7.11%

Source: Goetzmann, Ibbotson, and Peng (2001).

and vice versa. Empirical tests of this model have yielded mixed results. The problem with searching for predictability of stock returns over long horizons is that we have few independent observations from capital market history. The evidence using the NYSE data from 1815 finds some predicatability in subperiods of U.S. financial history but not over the sample as a whole. In recent reviews of long-horizon stock return predictability, Ang and Beckaert (2007) find only short-horizon stock return predictability, and Timmermann (2007) finds that any evidence of predictability is quite limited and short-lived.[7]

A New Approach: The Recovery Theorem

The problem with using historical data to estimate expected returns is that this is necessarily a backward-looking exercise. Indeed Ilmanen (2011) argues that historical average returns are particularly misleading measures of prospective long-term returns if expected returns vary over time and the past sample includes significant repricing. It should be possible, at least in principle, to gain a measure of the market's expectation of future returns from the prices of options and other derivative securities whose payoffs will depend on what happens to the market. As we shall see in Chapter 23, it is not so simple. The price of these securities will depend not only on the probability of future market movements good and bad but also to an extent on how risk averse the market is in valuing these securities. In new work, Stephen Ross (2011) shows how to disentangle the probability of future market movements from the degree of risk aversion in the market. In this way we

Table 10.3

Summary Statistics for U.S. Stocks, Bonds, Bills, and Inflation 1926–2011

	Arithmetic Mean (%)	Geometric Mean (%)	Standard Deviation (%)
Stocks	11.8	9.8	20.3
LT govt. bonds	6.1	5.7	8.4
T-bills	3.6	3.6	3.1
Inflation	3.1	3.0	4.2

Source: *Stocks, Bonds, Bills and Inflation*, 2005 Yearbook, Ibbotson Associates, Chicago.

[7]In this context, Brown (2007) oberves that this evidence, such as it is, is neither necessary nor sufficient for the existence of profitable trading strategies based on this predictability.

can obtain a market-based forward-looking estimate of future expected returns. While the details are rather technical and can be found in the appendix to this chapter, at any point of time it is then just a data processing exercise to recover the market probabilities from the prices at which derivative securities trade. Ross (2011) refers to this result as the "Recovery Theorem."

FORECASTING INDIVIDUAL SECURITY RETURNS

An individual security's expected returns are almost always based on estimates provided by analysts. The techniques used for obtaining these forecasts are contained in Chapters 17, 18, and 19, which discuss valuation models, earnings estimation, and efficient markets. In this section, we discuss some characteristics of these forecasts that need to be taken into account when forming portfolios.

Researchers have found that forecasts of analysts across the stocks they follow tend to be too optimistic and too diverse, having too high a mean and too much dispersion. Nevertheless, researchers have found that analysts' estimates do have information content [see, e.g., Elton, Gruber, and Grossman (1986)]. However, there is substantial error. A useful way to think about the information is that these are nuggets of gold in a large pile of rock. If this information is used directly as input into a portfolio optimizer, then the extreme estimates will result in a portfolio that includes very few securities, frequently heavily concentrated in only one or two. These heavily concentrated portfolios will have high risk. Given the substantial error in forecasts of expected return, the extra return from these portfolios is likely to be small, and given the higher risk, the portfolios are likely to perform poorly. The object then is to devise techniques that still utilize the information in the forecasts but result in well-diversified portfolios.

Diversification serves three purposes. First, diversified portfolios have lower risk than more concentrated portfolios selected from the set of diverse forecasts. Second, it is generally believed that analysts' estimates have some information content but with lots of random noise. If the errors are uncorrelated, then a larger portfolio reduces the amount of random noise and increases the chance that the extra return is observed. Third, increasing the number of securities and reducing the amount invested in any single one reduces the amount invested in a security due to the extreme estimate of one analyst.

The easiest way to ensure diversification is to put upper limits on the amount invested in each security. A 2% upper limit will guarantee at least 50 securities in any given portfolio. A 1% limit would guarantee 100 securities. Upper limits are useful and are a common feature in most analysis. The difficulty is that many securities will be at the upper limit. If we believe securities with high forecasted expected returns are more desirable on average, then we would like to hold these securities in higher proportions. An upper limit of 1% to ensure at least 100 securities in the portfolio may be harmful if some of the securities with highest forecasted expected returns or desirable risk characteristics are securities in which we would like to invest more heavily. How else can one ensure reasonable diversification while allowing higher allocations to some securities?

One way to do this is to allow higher upper bounds but to process analysts' data to reduce some of the extreme variability. The simplest way to make forecasts less extreme and avoid the difficulties caused by this is to move all the forecasts part way to the mean and adjust the whole distribution of analysts' estimates so that it has a mean consistent with what we believe is appropriate for the type of securities being examined.

For example, if we employ analysts who forecast a 16% return for the average equity security and we forecast a market return of 12% for equities, we can first lower

all analysts' individual estimates by 4%.[8] Then, to get rid of the extreme forecasts, we can adjust all forecasts toward the mean. For example, if an analyst's forecast is an expected return of 20% for stock ABC, we would first adjust the forecasts so all of the forecasts have a mean consistent with our beliefs or, in this case, reduce it by 4% to 16%. Then we further adjust it by some percentage (e.g., 50%) of the difference of the forecast from the forecasted mean. Thus the forecast would be $16 - (1/2)\ 4 = 14$ for ABC. This type of adjustment preserves the rank order of the forecasts and, by making them less extreme, results in a more diversified portfolio. The difficulty with this simple adjustment is that if one believes that the securities differ in risk, the simple adjustment does not preserve the rank order of what analysts believe are good purchases (e.g., an expected return more than commensurate with their risk).

If one believes that an equilibrium model describes reality, then to maintain a relationship between what the analyst believes is the extra return above what is required, before and after the adjustment, one should adjust deviations toward what the security should return in equilibrium rather than toward the mean.

For example, assume the CAPM, which is shown in Chapter 13, assumes a linear relationship between expected return and beta. Then, if we were to plot the security analyst's estimate of expected return versus beta and fitted a line, we would get a plot as shown in Figure 10.1. This is called the *empirical CAPM*. If all securities are plotted along the empirical CAPM, then by using analysis discussed in Chapter 9 on simple rules, it can be shown that all securities would be held in market proportions, and none of the information in the analysts' forecasts would be utilized. Define the distance between the analysts' estimates of expected return and the expected return of the empirical CAPM as alpha (α). The normal adjustment is to lower positive alpha and raise negative alpha so that there is a closer clustering around the line. The closer the securities plot to the security market line, the more diversified the portfolio. For example, by cutting all alphas in half, the optimizer will produce a less extreme and more diversified portfolio.

If the manager believes that an equilibrium model is appropriate, this also preserves the information in the analysts' forecasts concerning whether the security gives above- or below-equilibrium return. A simple adjustment to the mean, as discussed earlier, could cause a security with a high beta and positive alpha to end up with a negative alpha.

Using an adjustment to an equilibrium model preserves the sign and rank order of the alphas. After adjusting the individual alphas, the empirical security market line can be

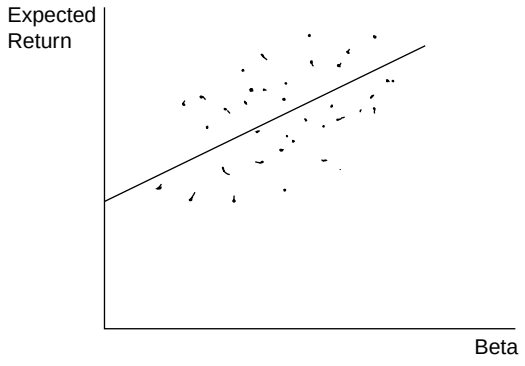


Figure 10.1 Relationship between expected return and beta.

[8]Because the market is a value-weighted average, we need to value-weight the analysts' forecasts to determine how much we need to adjust the mean.

lowered to match the belief concerning the expected return for the market as a whole. The forecast for an individual security is its alpha plus the expected return given its beta from the security market line adjusted as just described. While we present this procedure using the CAPM as an example of an equilibrium model, any of the equilibrium models discussed in Chapters 12–15 could be used to produce similar results.

Another adjustment that can be made to render analysts' expected return forecasts more usable is to recognize that different analysts and different sources of forecasts have different information content. Common sense would suggest that we should adjust the forecasts of less-accurate forecasters more than those of forecasters who make more accurate predictions.

If we have single forecasts for each security but multiple forecasters, then there is no special way to determine how much to adjust the forecasts of different analysts, except the principle of adjusting the least accurate more. If there are multiple forecasts for the same security, then there are procedures for determining the optimum weight of each forecast. The details are beyond this discussion, but the interested reader can pursue the references in the footnotes.[9] Security estimates for bonds are much less extreme and usually are not biased upward. Thus the techniques discussed in Chapters 21 and 22 can be used directly.

Up to now, we have assumed that analysts estimate expected returns. However, in many firms, analysts simply provide a discrete rank for each security. We now turn to a discussion of handling this type of data.

PORTFOLIO ANALYSIS WITH DISCRETE DATA

Often analysts' information about expected return comes in the form of discrete rankings rather than an estimate of expected return. For example, one common ranking used by industry is to place a stock in one of the following five categories:

1. strong buy
2. buy
3. hold
4. sell
5. strong sell

If this is the form of analyst information, then different techniques for forming portfolios are required.

The optimum way to utilize these data depends on how one believes the groups were formed in the first place. In most cases, the belief is that they were formed without any consideration of the risk characteristics of the securities. In this case, there is no single optimum method for utilizing these data. However, there are a number of methods that are sensible.

One technique that can be used is to construct an index fund out of the top group or groups. To construct an index fund, one would decide on the return-generating process that best fits the data (see Chapters 7 and 8) and then determine the sensitivities of the market to the factors in the model. Once these are determined, one would construct a portfolio from the top-ranked securities with the same sensitivity as the market to each of the factors and that has minimal residual risk. Such a portfolio has some nice characteristics. First, if the rankings contain no information, then one has constructed a portfolio that should mimic an index fund. Second, if there is information in the rankings, then the portfolio should have

[9]For measurements of forecasts, see Chapter 27. For combining forecasts differently, see Figlewski (1983) and Clemen (1989) and references therein.

volatility similar to the market and be highly correlated with the market (move up and down with the market) but have extra return. In other words, such a portfolio would perform like an enhanced return index fund. The only condition under which the portfolio would not perform well is if the information in the rankings were perverse, that is, the highest-ranked securities were actually the worst securities to hold.

Alternatively, one could construct a minimum-risk portfolio out of the top group or groups. Why is this sensible? If one has no basis to differentiate among the securities in the top group with respect to expected return, then one should assign them all the same expected return. If they all have the same expected return, then so does any linear combination. If all portfolios have the same expected returns, then the optimum course of action is to find a portfolio of the top-ranked securities that have minimum risk. This is obtained by solving a simple quadratic programming problem.

What can be done if you believe the groups are formed by expected return but you are unwilling to make any estimate of the risk of the securities? In this case, it can be shown that the optimum strategy is to hold each security in a group in the same proportion. All groups one believes have an expected return above the riskless rate should be held. The proportion invested in each group is proportional to that group's excess return (expected return above the riskless rate). If X_1 and X_2 are the amounts to invest in groups 1 and 2, R_1 and R_2 are their expected returns, and R_F is the riskless rate, then

$$\frac{X_1}{X_2} = \frac{\bar{R}_1 - R_F}{\bar{R}_2 - R_F}$$

The estimates of $\bar{R}_1$ and $\bar{R}_2$ are, of course, made by the portfolio's manager, because the data do not provide them directly.

If one believes that the groupings were based on both risk and return, then the optimum way to utilize the data changes. There are a number of different ways that analysts could form groups. One possibility is by ranking by excess return to beta.[10] In this case, it can be shown that the optimum portfolio consists of holding the first group in its entirety. The amount to invest in each stock in the group is inversely related to the residual risk.

APPENDIX
THE ROSS RECOVERY THEOREM—A NEW APPROACH TO USING MARKET DATA TO CALCULATE EXPECTED RETURN

Up until this point, most practitioners have been limited to the use of historical data to estimate expected returns and measures of risk as inputs to the portfolio problem. The difficulty is that such measures are inherently backward looking, while the appropriate measures of expected return and risk should be forward looking. There is a general understanding that derivative markets give us considerable insight into what market participants think will happen in the future. Indeed, up until Black and Scholes (1973), many practitioners thought that options were a "fair game" in the sense that option values should reflect the expected value of anticipated future payoffs from those options. If this is true, it should be possible at least in principle to infer the probability distribution of stock

[10]Proofs are contained in Elton and Gruber (1987). The article contains the solution for a number of other risk assumptions.

returns in the future from the prices at which options are trading in the option markets. If you could recover probabilities from option values, you could derive forward-looking estimates of expected value. Unfortunately, the prices of deep out of the money options in most cases exceed this actuarily determined formula, which led many to question the rationality of the option markets. Since Black and Scholes (1973), we now understand that option values reflect not only the probability of future events but also the risk aversion of investors who buy and sell these options. Disentangling the probability measure from the degree of risk aversion in the markets seems to be an impossible challenge.

In a new paper Stephen Ross (2011) shows how to resolve this problem and recover the underlying market probabilities implied in option prices. In a very simple world where there are only two possible market return outcomes, R_h and R_i, it is always possible to find a portfolio of assets that pay off under these two contingencies that has no risk and thus earns the risk-free return R_f:

$$q(1+R_h)+(1-q)(1+R_l)=1+R_f, \text{ where } q=\frac{R_f-R_l}{R_h-R_l} \tag{A.1}$$

Using the weights q and $(1 - q)$ we can then value any derivative security whose payoff depends only on R_h and R_l. This is referred to as the binomial option pricing formula and is discussed in Chapter 23. Because a weighted average of the possible values of the security next period gives the value of the security when discounted back at the risk-free rate R_f, the weights q and $(1 - q)$ are often referred to as "risk-neutral probabilities." The analysis can be extended to more than two possible outcomes, even to the case where there is a continuous range of possible returns. In a simple three-outcome example we could think of the risk-neutral probability going from a low value to a medium or high value, or starting out at a medium value and rising or falling in value, or starting high and falling in value. As long as there are a sufficient number and range of derivative securities trading, we can essentially observe the entire set of risk-neutral probabilities, where q_{ij} is the risk-neutral probability that a security currently trading at S_i could take any one of a number of values S_J in the next period of time. However, it is not immediately obvious how these risk-neutral probabilities relate to the actual probability distribution of returns in the next period. We need to obtain estimates of expected value and risk as inputs to the portfolio problem.

The contribution of Ross (2011) is to relate these risk-neutral probabilities to the actual probability distribution of returns through the utility function of the representative market individual. If such an individual is willing to pay \$1 for a security that pays off $\$(1 + R_h)$ in good times and $\$(1 + R_l)$ in bad times, then it must be true that

$$p\times m_h(1+R_h)+(1-p)\times m_l(1+R_l)=1 \tag{A.2}$$

where m_h is the discount factor appropriate for good times, m_l is the discount factor appropriate for bad times, and p is the probability that good times will occur. Clearly, comparing the two equations, the discount factors are given by

$$m_h=\frac{q}{p(1+R_f)} \quad \text{and} \quad m_l=\frac{1-q}{(1-p)(1+R_f)} \tag{A.3}$$

which shows that the spread between p and q is a measure of the risk aversion implicit in the discount factors associated with returns in the future. These discount factors depend on whether good times or bad times occur in the future and are not known at the time the

investment is made. For this reason, they are referred to as "stochastic discount factors." This result (0.3) can be generalized to more than two possible states of the economy, so that $m_{ij} = q_{ij} / P_{ij}(1 + R_f)$.

How can we interpret these stochastic discount factors? In the appendix to Chapter 13, we see that Equation (A.2) flows naturally from the first-order conditions where the investor is solving a multiperiod consumption and investment problem. In that context

$$m_{ij} = \delta \frac{U'(c_j)}{U'(c_i)} \tag{A.4}$$

where $U'(c_i)$ and $U'(c_j)$ are the marginal utility of consumption when the economy is in states i and j, respectively, and δ is the investor's personal discount rate. We can think of these marginal utilities as those of the representative investor. For this reason, the stochastic discount factor is sometimes referred to as an intertemporal marginal rate of substitution. Substituting this expression in for the stochastic discount factor, we have immediately that

$$U'(c_i)q_{ij} = \delta(1+R_f)p_{ij}U'(c_j) \tag{A.5}$$

If there are three possible states of the economy, bad, normal, and good (l, m, and h), then we can express Equation (A.5) in a straightforward matrix equation:

$$\begin{bmatrix} U'(c_l) & 0 & 0 \\ 0 & U'(c_m) & 0 \\ 0 & 0 & U'(c_h) \end{bmatrix} \times \begin{bmatrix} q_{ll} & q_{lm} & q_{lh} \\ q_{ml} & q_{mm} & q_{mh} \\ q_{hl} & q_{hm} & q_{hh} \end{bmatrix} = \delta\left(1+R_f\right) \begin{bmatrix} p_{ll} & p_{lm} & p_{lh} \\ p_{ml} & p_{mm} & p_{mh} \\ p_{hl} & p_{hm} & p_{hh} \end{bmatrix} \times \begin{bmatrix} U'(c_l) & 0 & 0 \\ 0 & U'(c_m) & 0 \\ 0 & 0 & U'(c_h) \end{bmatrix}$$

or in a more compact and more general expression,

$$DQ = \delta(1+R_f)PD \tag{A.6}$$

where D is a diagonal matrix with marginal utilities on the diagonal and Q is a matrix of all of the risk-neutral probabilities inferred from the prices of derivative securities. P is the matrix of probabilities we are trying to infer, where in the three-state example, the first row gives the probabilities that the economy will stay in the bad state or will move to the normal or to the good state, respectively, the second row gives the same probabilities starting in the normal state, while the last row starts in the good state. Each row of this matrix adds up to 1, which in matrix terms implies that $Pi = i$, where i is a vector of ones. We can solve this expression for P to obtain

$$P = \frac{1}{\delta(1+R_f)} DQD^{-1} \tag{A.7}$$

where D^{-1} is a diagonal matrix with the reciprocal of the marginal utilities on the diagonal. But how can we solve this expression for P when we do not know the marginal utilities in the diagonal matrix D? The trick here is to rearrange the expression, premultiplying both sides of the equation by D^{-1} and postmultiplying both sides of the equation by the vector of ones t to obtain $QD^{-1}t = \delta(1 + R_f)D^{-1}Pt$. Using the result that $Pt = t$, we finally obtain the equation

$$QY = \lambda y \tag{A.8}$$

where γ is a vector containing the reciprocals of the marginal utilities and $\lambda = \delta(1 + R_f)$. The reader will note that Equation (A.8) is a standard eigenvalue problem. Ross (2011) shows that it has a unique solution under quite general conditions, so we can actually solve for both the marginal utilities (given as the reciprocal of elements in the solution vector γ) and the personal discount factor and risk-free rate once we know the matrix of risk-neutral probabilities we can observe from the prices at which derivative securities trade. Once we know the marginal utilities, we just substitute them into Equation (A.7) to recover all relevant market probabilities. Ross (2011) refers to the notion that one can essentially observe market probabilities from the prices at which derivative securities trade as the *Recovery theorem*.

BIBLIOGRAPHY

1. Ang, Andrew, and Beckaert, Geert. "Stock Return Predictability: Is It There?" *Review of Financial Studies* **20,** No. 3 (2007), pp. 651–707.
2. Black, Fisher, and Litterman, Robert. "Global Portfolio Optimization," *Financial Analyst Journal*, **48** (1992), pp. 20–43.
3. Black, Fisher, and Scholes, Myron. "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, **81** (1973), pp. 637–654.
4. Brown, S. J. "Optimal Portfolio Choice under Uncertainty: A Bayesian Approach," in V. S. Bawa, S. J. Brown, and R. W. Klein (eds.), *Estimation Risk and Optimal Portfolio Choice* (Amsterdam: North Holland, 1979).
5. Brown, Stephen. "Elusive Return Predictability: Discussion," *International Journal of Forecasting* **24,** No. 1 (January–March 2008), pp. 19–21.
6. Campbell, John Y., and Shiller, Richard. "Stock Prices, Earnings and Expected Dividends," *Journal of Finance*, **43,** No. 3 (1988), pp. 661–676.
7. Clemen, Robert. "Combining Forecasts: A Review and Annotated Bibliography," *International Journal of Forecasting*, **5** (1989), pp. 559–583.
8. Comer, George. "Hybrid Mutual and Market Timing Performance," *Journal of Business*, **79** (2006), pp. 771–797.
9. DeMiguel, Victor, Garlappi, Lorenzo, and Uppal, Raman. "Optimal versus Naive Diversification: How Inefficient Is the 1/*N* Portfolio Strategy?" *Review of Financial Studies*, **22,** No. 5 (2009), pp. 1915–1953.
10. Dimson, Elroy, Marsh, Paul, and Staunton, Mike. *Triumph of the Optimists: 101 Years of Global Investment Returns* (Princeton: Princeton University Press, 2002).
11. Elton, E., and Gruber, M. "Portfolio Analysis with Partial Information: The Case of Grouped Data Management," *Science*, **33,** No. 10 (1987), pp. 1238–1246.
12. Elton, E., Gruber, M., and Padberg, Manfred W. "Simple Criteria for Optimal Portfolio Selection," *Journal of Finance*, **31,** No. 5 (1976), pp. 1341–1357.
13. Fama, Eugene, and French, Ken. "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, **22,** No. 1 (1988), pp. 3–25.
14. Fama, Eugene, and French, Ken. "Business Conditions and Expected Returns," *On Stocks and Bonds*, **25,** No. 1 (1989), pp. 23–49.
15. Farrell, James. *Guide to Portfolio Management* (New York: McGraw-Hill, 1997).
16. Figlewski, Steve. "Optimal Price Forecasting Using Survey Data," *Review of Economics and Statistics*, **65,** No. 1 (1983), pp. 27–38.
17. Goetzmann, William, and Jorion, Phillipe. "Testing the Predictive Power of Dividend Yields," *Journal of Finance* **48,** No. 2 (1993), pp. 663–679.
18. Goetzmann, William, and Jorion, Phillipe. "A Longer Look at Dividend Yields," *Journal of Business* **68,** No. 4 (1995), pp. 483–508.
19. Goetzmann, William, Ibbotson, Roger, and Peng, Liang. "A New Historical Database for the NYSE 1815 to 1925: Performance and Predictability," *Journal of Financial Markets*, **4,** No. 1 (2001), pp. 1–32.

20. Goetzmann, William, and Ibbotson, Roger. *The Equity Risk Premium, Essays and Explorations* (New York: Oxford University Press, 2006).
21. Henriksson, Roy D., and Merton, Robert C. "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Skills," *Journal of Business*, **54** (1981), pp. 513–553.
22. Ilmanen, Antti. *Expected Returns: An Investor's Guide to Harvesting Market Returns* (Chichester, UK: John Wiley, 2011).
23. Ibbotson Associates. "Stocks Bonds Bills and Inflations," *2005 Yearbook* (Chicago: Ibbotson Associates, 2005).
24. Jorion, Philippe. "Bayes-Stein Estimation for Portfolio Analysis," *Journal of Financial and Quantitative Analysis*, **21,** No. 3 (1986), pp. 279–296.
25. Jorion, Philippe. "Portfolio Optimization in Practice." *Financial Analysts Journal*, **48** (1992), pp. 68–74.
26. Kan, Raymond, and Guofu, Zhou. "Optimal Portfolio Choice with Parameter Uncertainty." *Journal of Financial and Quantitative Analysis*, **42,** No. 3 (2007), pp. 621–632.
27. Poterba, James, and Summers, Lawrence. "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, **22,** No. 1 (Oct. 1988), pp. 27–59.
28. Ross, Stephen A., "The Recovery Theorem." NBER working paper No. w17323 (2011). Available at SSRN: http://ssrn.com/abstract=1918653.
29. Rozeff, Michael. "Dividend Yields Are Equity Risk Premiums," *Journal of Portfolio Management*, **11,** No. 1 (Fall 1984), pp. 68–75.
30. Timmermann, Alan. "Elusive Return Predictability," *International Journal of Forecasting*, **24,** No. 1 (Jan.–March 2008), pp. 1–18.
31. Treynor, Jack, and Muzay, F. "Can Mutual Funds Outguess the Market?" *Harvard Business Review*, **44** (1966), pp. 131–136.
32. Tu, Jun, and Guofu, Zhou. "Markowitz Meets Talmud: A Combination of Sophisticated and Naive Diversification Strategies." *Journal of Financial Economics*, **99,** No. 1 (2011), pp. 204–215.

11

How to Select among the Portfolios in the Opportunity Set

In Chapter 1, we pointed out that to solve any decision problem, one needed to define an opportunity set and a way to pick the optimum portfolio from the opportunity set. The subject of earlier chapters was how to obtain an opportunity set. The subject of this chapter is picking the optimum portfolio. In what follows, we discuss various techniques that have been proposed for selecting the optimum portfolio.

CHOOSING DIRECTLY

The simplest way to select among portfolios in the opportunity set is to directly compare them. Many investment professionals and academics are skeptical concerning the investor's ability to specify the trade-offs necessary to implement more formal procedures for making these choices.

Consider the three portfolios shown in Table 11.1. These portfolios are associated with an efficient frontier assuming riskless lending and borrowing. The Tangency Portfolio has an expected return of 10 and a standard deviation of 10, and the risk-free rate is 4%.

How can an investor directly choose among these portfolios? Normally, investors don't think in terms of expected return and standard deviation of return so that the investor or her advisor often expresses the choice in terms of the likelihood of outcomes that might be important to the investor. Alternatively, the advisor can present the investor with probability distributions representing the payoff for various alternatives.

First, consider expressing the choice in terms of returns an investor cares about. Most investors are concerned with negative outcomes. One way to determine the probability of a negative outcome is as follows. Assume returns are normally distributed. The mean return

Table 11.1 Return and Risk on Portfolios in the Efficient Set

	Amount Invested			
Portfolio	Tangency Portfolio	Riskless Asset	$\bar{R}$	σ
1	1/2	1/2	7	10
2	3/4	1/4	8.5	15
3	1	0	10	20

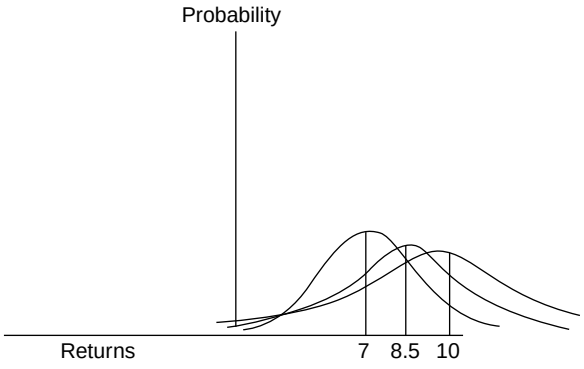


Figure 11.1 The distribution of returns for portfolios shown in Table 11.1

of portfolio 1 in Table 11.1 is 7; the mean is $7/10 = 0.7$ standard deviations from zero using a normal table. A standard deviation of 0.7 from the mean occurs 31% of the time. Thus one way to express the choices shown in Table 11.1 is as follows. Which do you prefer, an investment that on average returns 10% but 31% of the time has returns below zero, or an investment that returns 7% on average and has a 24% chance of negative returns? Once this choice has been made, the advisor can select other possible portfolios to compare to the preferred choice and in this manner narrow the choice to a portfolio on the efficient frontier. Alternatively, the advisor or investor can draw the distribution of outcomes for the portfolio in question, and the investor can examine the distributions and select the preferred choice. The distributions for portfolios 1, 2, and 3 are shown in Figure 11.1. Although this is not high technology, it may well result in the best choices. We now examine more formal procedures for selecting the optimum portfolio.

AN INTRODUCTION TO PREFERENCE FUNCTIONS

We start our formal discussion of the choice between risky assets with a simple example. Consider the two alternatives shown in Table 11.2. Investment A and investment B each have three possible outcomes, each equally likely. Investment A has less variability in its outcomes but has a lower average outcome.

One approach to choosing between them is to specify how much more valuable the large outcomes are relative to the small outcomes and then to weight the outcomes by their value and find the expected value of these weighted outcomes. The idea of adding up or averaging weighted outcomes is very common. Consider, for example, how the winning team is selected in hockey. Table 11.3 shows the hypothetical records for two hockey teams.

Table 11.2 Two Alternative Investments

Investment A		Investment B	
Outcome	Probability of Outcome	Outcome	Probability of Outcome
15	1/3	20	1/3
10	1/3	12	1/3
5	1/3	4	1/3

Table 11.3 Data for Ranking Hockey Teams

	Islanders	Flyers
Wins	40	45
Ties	20	5
Losses	10	20

Current practice weights wins by two, ties by one, and losses by zero. With this weighting scheme, the Islanders would be leading the Flyers 100 to 95. But there is nothing special about this weighting scheme. A league interested in deemphasizing the incentive for ties might weight wins by four, ties by one, and losses by zero. In this case, the Flyers would be considered the dominant team, 185 to 180. If we denote W as the result (win, tie, lose), $U(W)$ as the value of this result, and $N(W)$ as the number of times (games) that W occurs, then to determine the better team, we calculate

$$\sum_W U(W)N(W)$$

The team with the higher U is considered the better team. For example, utilizing current practice, U (win) = 2, U (tie) = 1, and U (loss) = 0. Applying the formula to the Islanders yields

$$U = 2(40) + 1(20) + 0(10) = 100$$

This is the 100 we referred to earlier. While the particular function $U(W)$ differs between situations, the principle is the same. Traditionally, instead of using the number of outcomes of a particular type, the proportion is used. There were 70 hockey games in our example. If $P(W)$ is the proportion of the total games that resulted in outcome W, then $P(W) = N(W)/70$. Dividing through by 70 will not affect the ordering of *teams*. Weighting a function by the proportion of each outcome is equivalent to calculating an average or expected value. Letting $E(U)$ designate the expected value of U yields[1]

$$E(U) = \sum_W U(W)P(W)$$

When we apply this principle to the decision problem shown in Table 11.2, we have special names for the principle. The weighting function is called a *utility function* and the principle is called the *expected utility theorem.* Consider the example shown in Table 11.2 and a set of weights as shown in Table 11.4.

Table 11.4 A Weighing Function

Outcome	Weight	Value of Outcome
20	0.9	18
15	1.0	15
12	1.1	13.2
10	1.2	12
5	1.4	7
4	1.5	6

[1]Should be read as the sum over all results.

We have called the last column in the table the value of the outcome. Alternatively, it could be called the utility of an outcome. If this was the weighting function the investor felt was appropriate, then she would compare the expected utility of investments A and B using this function. For example, the expected utility of A is

$$U(15)(1/3) + U(10)(1/3) + U(5)(1/3)$$

Referring to the weighting function, we have

$$15(1/3) + 12(1/3) + 7(1/3) = 34/3$$

and the expected utility of investment B is

$$U(20)(1/3) + U(12)(1/3) + U(4)(1/3) = 18(1/3) + (13.2)(1/3) + 6(1/3) = 37.2/3$$

In this situation, the investor would select investment B because it offers the higher average or expected utility. In general, we can say that the investor will choose among alternatives by maximizing expected utility or maximizing

$$E(U) = \sum_{W} U(W)P(W)$$

Consider a second example. Table 11.5 lists three separate investments. Assume the investor has the following utility function:

$$U(W) = 4W - (1/10)W^2$$

Then the utility of 20 is $80 - (1/10)(400) = 40$; the utility of 18 is $72 - (1/10)(324) = 39.6$; and the utility of 14 is $56 - (1/10)(196) = 36.4$.

The rest of the values are shown in Table 11.6. The expected utility of the three investments is found by multiplying the probability of each outcome times the value of the outcome:

$$\text{Expected utility A} = (40)(3/15) + (39.6)(5/15) + (36.4)(4/15) + (30)(2/15) + 1\ (20.4)(1/15) = 544/15 = 36.3$$

$$\text{Expected utility B} = (39.9)(1/5) + (30)(2/5) + (17.5)(2/5) = 134.9/5 = 26.98$$

$$\text{Expected utility C} = (39.6)(1/4) + (38.4)(1/4) + (33.6)(1/4) + (25.6)(1/4) = 137.2/4 = 34.4$$

Thus an investor with the utility function discussed earlier would select investment A.

Table 11.5 Outcomes and Associated Probabilities for Three Investments

Investment A		Investment B		Investment C	
Outcome	Probability	Outcome	Probability	Outcome	Probability
20	3/15	19	1/5	18	1/4
18	5/15	10	2/5	16	1/4
14	4/15	5	2/5	12	1/4
10	2/15			8	1/4
6	1/15				

Table 11.6 Including Utility

Investment A			Investment B			Investment C		
Outcome	Utility of Outcome	Probability	Outcome	Utility of Outcome	Probability	Outcome	Utility of Outcome	Probability
20	40	3/15	19	39.9	1/5	18	39.6	1/4
18	39.6	5/15	10	30	2/5	16	38.4	1/4
14	36.4	4/15	5	17.5	2/5	12	33.6	1/4
10	30	2/15				8	25.6	1/4
6	20.4	1/15						

If the investor is consistent in her choices, then the choice of the preferred investment, using the expected utility theorem, is identical to the choice made by examining the investment directly. Note that the weighting function in Table 11.4 values small outcomes more heavily than large outcomes. Most investors prefer more wealth to less wealth and would prefer money with certainty rather than engage in a gamble with the same expected value. These types of observations about investor behavior allow us to place restrictions on what are appropriate utility functions. This is discussed in the appendix. However, even taking these properties into account, the number of potential utility functions is enormous. It follows that, in having an investor make choices between a series of simple investments, we can attempt to determine the weighting (utility) function that the investor is implicitly using. Applying this weighting function to more complicated investments, we should be able to determine which one the investor would choose.

A number of brokerage firms and banks have developed programs to extract the utility function of investors by confronting them with a choice between a series of simple investments. These have not been particularly successful. Many investors are not consistent when faced with a series of choice situations. Also, many investors, when faced with more complicated choice situations, encounter aspects of the problem that were not of concern to them in the simple choice situations. This has led to an alternative way of analyzing the problem.

RISK TOLERANCE FUNCTIONS

Note that the portfolio problem is expressed as a choice between mean returns and standard deviation of return. Thus any utility function can alternately be expressed the same way. This has resulted in a proposal to express expected utility maximization as maximizing

$$f = \overline{R} - \frac{\sigma^2}{T}$$

where T is referred to as *risk tolerance* and expresses the investor's trade-off between expected return and variance of return. The higher T, the "more tolerant" the investor is of risk and the higher the risk of the portfolio selected. Table 11.7 shows the choice for two investors: investor A, with a risk tolerance of 100, and investor B, with a risk tolerance of 150. Their choices are applied to the investment problem shown in Table 11.1.

With these choices and risk tolerances, investor A would select investment 2 and investor B would select investment 3. One way to apply the risk tolerance idea is to simply use it to evaluate the investments being considered. When we assume riskless lending and borrowing, the optimum proportion to invest in the Tangency Portfolio (X_T) and the

Table 11.7 Choices Using Risk Tolerance

			Value of Investments to Investors A and B	
	$\bar{R}$	σ	A	B
1	7	10	6	6 1/3
2	8.5	15	6.25	7
3	10	20	6	7 1/3

amount to lend or borrow $(1 - X_T)$ can be determined directly. Using the preceding equation and substituting in the formula for the expected return and variance of the portfolio of debt and stock, finding the value of X_T that maximizes the function yields[2]

$$x_T = \frac{T}{2}\left(\frac{\bar{R}_T - R_F}{\sigma_T^2}\right)$$

For the example discussed in Table 11.1, the Tangency Portfolio had a mean return of 10 and a standard deviation of 20, and the riskless rate was 4%. Thus, for investor A, with a risk tolerance of 100, we have

$$x_T = \frac{100}{2}\left[\frac{10-4}{400}\right] = \frac{6}{8} = \frac{3}{4}$$

And for investor B, with a risk tolerance of 150, we have

$$x_T = \frac{150}{2}\left[\frac{10-4}{400}\right] = 1\frac{1}{8}$$

Once again, to implement this, one needs to estimate an investor's risk tolerance. Risk tolerance is easier to obtain from an investor because it is a single number. In implementing utility functions, one has to determine both the functional form of the investor's utility function and the parameters. Although we can specify some general characteristics of utility functions,

---

[2]With riskless lending and borrowing,

$$\bar{R}_p = x_T\bar{R}_T + (1 - x_T)R_F = R_F + x_T(\bar{R}_T - R_F)$$

$$\text{and } \sigma_p^2 = x_T^2\sigma_T^2$$

where

x_T is the proportion in the tangent portfolio

$\bar{R}_T$ and σ_T^2 refer to the mean return and standard deviation of the tangent portfolio

The risk tolerance function substituting in the mean and standard deviation given earlier is

$$f = \left[R_T + x_T\left(\bar{R}_T - R_F\right)\right] - \frac{x_T^2\sigma_T^2}{T}$$

The derivative is set to zero:

$$\frac{df}{dx_T} = \bar{R}_T - R_F - \frac{2}{T}x_T\sigma_T^2 = 0$$

Solving for x_T gives the expression in the text.

lots of functional forms are reasonable. As discussed, it is hard to get investors to answer choice situations consistently enough to be able to determine the functional form and the parameters of the function. Thus most firms that are trying to determine a specific portfolio for a client rely on the risk tolerance framework and devise questionnaires to determine a reasonable risk tolerance for an investor. Alternatively, one can simply ask the investor how much she would put in the Tangency Portfolio and solve for the investor's risk tolerance. In the preceding example, if an investor desired to invest three-fourths of her wealth in the Tangency Portfolio, then one could use this to solve for X_T and get 100.

SAFETY FIRST

A second alternative to the expected utility theorem that is advocated by many is a group of criteria called *safety-first models*. The origin of these models is a belief that decision makers are unable, or unwilling, to go through the mathematics of the expected utility theorem but rather will use a simpler decision model that concentrates on bad outcomes. The name "safety first" comes about because of the emphasis each of the criteria places on limiting the risk of bad outcomes. Three different safety-first criteria have been put forth. The first, developed by Roy (1952), states that the best portfolio is the one that has the smallest probability of producing a return below some specified level. If R_P is the return on the portfolio and R_L is the level below which the investor does not wish returns to fall, Roy's criterion is

$$\text{minimize Prob } (R_P < R_L)$$

If returns are normally distributed, then the optimum portfolio would be the one where R_L was the maximum number of standard deviations away from the mean. For example, consider the three portfolios shown in Table 11.8. Assume 5% is the minimum return the investor desires. The investor wishes to minimize the chance of getting a return below 5%. If the investor selects portfolio A, then 5% is 1 standard deviation below the mean. The chance of getting a return below 5% is the probability of obtaining a return more than 1 standard deviation below the mean. If the investor selects investment B, then 5% is 2.25 standard deviations below the mean. The probability of obtaining a return below 5% is the probability of obtaining a return more than 2.25 standard deviations below the mean. If he selects investment C, the probability of obtaining a return below 5% is the probability of obtaining a return more than 1.5 standard deviations below the mean. Because the odds of obtaining a return more than 2.25 standard deviations below the mean are less than the odds of obtaining a return more than 1.5 or 1 standard deviation less than the mean, investment B is to be preferred.

As a second example, return to the problem shown in Table 11.1. Assume the investor wants to avoid negative outcomes. Portfolio A has a mean that is 7/10 or 0.7 standard deviations above zero, B has a mean that is 8.5/15 or 0.57 standard deviations above zero, and

Table 11.8 Mean Returns, Standard Deviations, and Lower Limits

	Portfolio		
	A	B	C
Mean return	10	14	17
Standard deviation(s)	5	4	8
Difference from 5%	-1σ	-2.25σ	-1.5σ

C has a mean that is 10/20 or 0.5 standard deviations from zero. Thus A has the lowest probability of returns below zero and is preferred using Roy's criterion.

To determine how many standard deviations R_L lies below the mean, we calculated R_L minus the mean return divided by the standard deviation. To satisfy Roy's criterion, if returns are normally distributed, we

$$\text{minimize } \frac{R_L - R_P}{\sigma_P}$$

This is equivalent to maximizing minus this ratio, or

$$\text{maximize } \frac{\bar{R}_P - R_L}{\sigma_P}$$

This criterion should look familiar. If R_L were replaced by R_F, the riskless rate of interest, this would be the criterion we used throughout much of the book. All portfolios that are equally desirable under Roy's criterion would have the same value for this ratio. That is, they could be described by the following expression:

$$\frac{\bar{R}_P - R_L}{\sigma_P} = K$$

Furthermore, if K were larger, the portfolio would be more desirable under Roy's criterion. Rearranging this expression yields

$$\bar{R}_P = R_L + K\sigma_P$$

This is the equation of a straight line with an intercept of R_L and a slope of K. Thus all points of equal desirability (i.e., constant K) plot on a straight line, and the preferred line is one with the highest slope. This is shown in Figure 11.2, where the Ks are ordered such

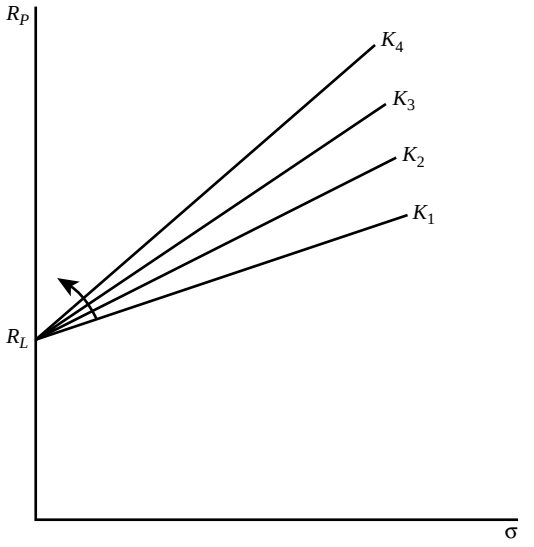


Figure 11.2 Lines of constant preference—Roy's criterion.

that $K_4 > K_3 > K_2 > K_1$. The Roy criterion with normally distributed returns produces a decision problem of exactly the same form as the portfolio problem with riskless lending and borrowing. In this case, R_L serves the role of the riskless rate, R_F. The desired portfolio is the feasible portfolio lying on the line in the most counterclockwise direction and is easy to find utilizing the standard techniques discussed earlier. Notice that the portfolio that maximizes Roy's criterion must lie along the efficient frontier in mean standard deviation space.

Although the analysis was performed assuming normally distributed returns, a similar result holds for any distribution that has first and second moments. The very same maximization problem follows from the use of Tchebyshev's inequality.[3]

The Tchebyshev inequality makes very weak assumptions about the underlying distribution. It gives an expression that allows the determination of the maximum odds of

[3]One of the ways to determine the probability of some outcome is the use of Tchebyshev's inequality. Tchebyshev's inequality allows one to determine the maximum probability of obtaining an outcome less than some value. It does not assume any distribution for returns. If a distribution was assumed, a more precise statement about probability could be made. Rather, it is a general statement applicable for all distributions.

The Tchebyshev inequality is

$$\text{Prob}\left(\left|\frac{R-\bar{R}_P}{\sigma_P}\right| > K\right) \leq \frac{1}{K^2}$$

where

R is the outcome

$\bar{R}_P$ is the mean return

σ_P is the standard deviation of return

K is a constant deviation of return

Since we are interested in the case where the lower limit is less than $\bar{R}_P$, the returns we are interested in are those less than $\bar{R}_P$. Therefore, the term in the absolute value sign is negative. Noting this, we can write the term in the parentheses as

$$\frac{R-\bar{R}_P}{\sigma_P} < -K$$

and the expression as

$$\text{Prob}\left(\left|\frac{R-\bar{R}_P}{\sigma_P}\right| < -K\right) \leq \frac{1}{K^2} \tag{11.1}$$

We can express the lower limit in Roy's criterion as the number of standard deviations K lies below the mean, or

$$K = \frac{\bar{R}_p - R_L}{\sigma_p} \tag{11.2}$$

Since Tchebyshev's inequality holds for any value of K, we can substitute the expression for K shown in Equation (11.2) into the left-hand side of Equation (11.1). Doing so, and simplifying, yields

$$\text{Prob}\,(R < R_L) \leq \frac{1}{K^2}$$

Since this is precisely Roy's criterion, we want to maximize K or maximize Equation (11.2). But this is exactly what we did in the case of the normal distribution.

obtaining a return less than some number. The use of this inequality leads to the same maximization problem and the same analysis as previously discussed. Thus mean–variance analysis follows from the Roy safety-first criterion.

The second safety-first criterion was developed by Kataoka, who suggests the following criterion: maximize the lower limit subject to the constraint that the probability of a return less than, or equal to, the lower limit is not greater than some predetermined value. For example, maximize R_L subject to the constraint that the chance of a return below R_L is less than or equal to 5%. If α is the probability of a return below the lower limit, then in symbols, this is

$$\text{maximizing } R_L$$

$$\text{subject to Prob } (R_P < R_L) \leq \alpha$$

If returns are normally distributed, we can analyze this criterion in mean standard deviation space. Earlier, we noted that if returns are normally distributed, then the probability of obtaining returns below some number depends on the number of standard deviations below the mean that the number lies. Thus the odds of obtaining a return more than 3 standard deviations below the mean is 0.13%, while the odds of obtaining a return more than 2 standard deviations below the mean is 2.28%. As an example, set $\alpha = 0.05$. From any table of the normal distribution, we see that this is met as long as the lower limit is at least 1.65 standard deviations below the mean. With $\alpha = 0.05$, the constraint becomes

$$R_L \leq \overline{R}_P - 1.65\,\sigma_P$$

Because we want to make R_L as large as possible, this inequality can be written as an equality. Writing it as an equality and rearranging, we obtain for a constant R_L

$$\overline{R}_P \geq R_L + 1.65\,\sigma_P$$

This is the equation of a straight line. Because the intercept is R_L as R_L changes, the line shifts in a parallel fashion. Figure 11.3 illustrates this for various values of R_L. The objective is to

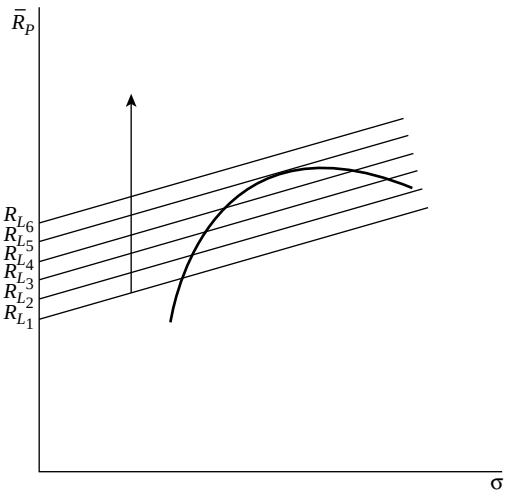


Figure 11.3 The portfolio choice problem with Kataoka's safety-first rule.

maximize R_L or to move as far up as possible (in the direction of the arrow). If there is no lending or borrowing, then a unique maximum exists, and it is the tangency point on the highest R_L line (R_{L5} in the example). Note that, as in the case of Roy's criterion, the optimum portfolio must be on the efficient frontier in mean standard deviation space. Once again, the same analysis follows if one chooses to use the Tchebyshev inequality rather than assuming normally distributed returns.

The final safety first criterion was put forth by Telser. He suggested that a reasonable criterion would be for an investor to maximize expected return, subject to the constraint that the probability of a return less than, or equal to, some predetermined limit was not greater than some predetermined number. In symbols, we have

$$\begin{aligned} &\text{maximize } \bar{R}_P \\ &\text{subject to Prob } (R_P \leq R_L) \leq \alpha \end{aligned}$$

Once again, it is convenient to rearrange the constraint. In the discussion of the Kataoka criterion, it was shown that if returns are normally distributed, this constraint becomes

$$R_L \leq \bar{R}_P - \text{constant } \alpha$$

Rearranging yields

$$\bar{R}_P \geq R_L + \text{constant } \alpha$$

In the last section, the constant was set equal to 1.65 for the example. In general, it depends on the value of α. As discussed earlier, when the equality holds, this expression is the equation of a straight line. Consider Figure 11.4. The efficient frontier and the constraint are plotted in that figure. All points above the line meet the constraint. In Figure 11.4 the feasible set is bounded by the straight line and the efficient frontier (the shaded area). In this case the optimum is point A. If the portfolio with the overall highest return lies above the line, it will be selected. If it does not, the constraint line excludes part of the efficient

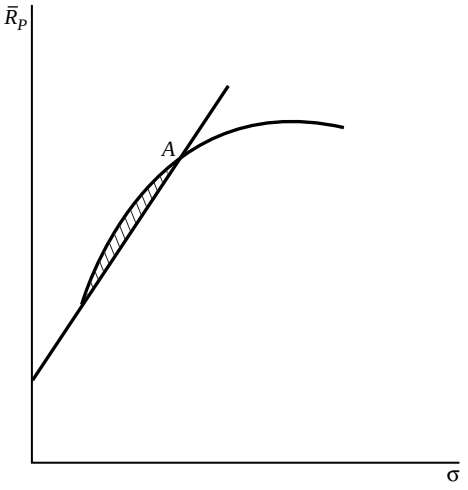


Figure 11.4 The investor's choice problem—Telser's criterion.

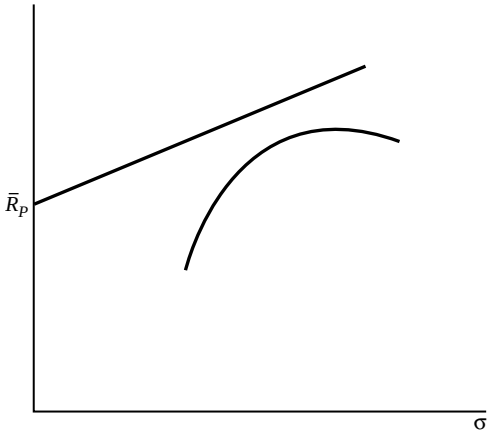


Figure 11.5 No feasible portfolio—Telser's criterion.

set. In this case, the feasible portfolio with the highest mean return will lie at the highest intersection of the efficient frontier and the constraint. In either case, the point selected will be on the efficient set. It is possible that there are no feasible points that meet the constraint. For example, in Figure 11.5, the constraint lies above the efficient set. In this case, there is no feasible portfolio lying above the constraint, and the criterion fails to select any portfolios. Note that with the Telser criterion, the optimum portfolio either lies on the efficient frontier in mean standard deviation space or it does not exist. As with the other two criteria, the same analysis follows if we use the Tchebyshev inequality rather than assuming normal returns.

Let us consider the portfolio selected by Kataoka and Telser. Consider the example in Table 11.1. For Kataoka, assume we want a probability of less than 5% of returns below the lower limit. Recall this means the lower limit is 1.65 standard deviations from the mean. For A this is $7-10\ (1.65) = -9.5$; for B we have $8.5-15(1.65) = -16.25$; and for C we have $10 - 20(1.65) = -23$. Thus A has the highest lower limit and is to be preferred. For Telser, assume the lower limit is -20%. Portfolios A and B have less than 5% chance of return below 20%, and B has the higher mean return; thus B is to be preferred.

The safety-first criteria were originally suggested as an appealing decision-making tool and an alternative to the expected utility framework of traditional analysis. We see in this section that, under reasonable sets of assumptions, they lead to mean–variance analysis and to the selection of a particular portfolio in the efficient set. With unlimited lending and borrowing at a riskless rate, the analysis may lead to infinite borrowing, an unreasonable prescription for managers. However, the difficulties lie not with the criteria but with the original assumption that investors can borrow unlimited amounts at a riskless rate of interest. Whether the safety-first criteria are reasonable criteria can be answered only by the readers themselves. To some, they seem sensible as a description of reality. To others, the fact that they may be inconsistent with expected utility maximization leads to their rejection. If one accepts one of the safety-first criteria and believes that the probability distribution of returns is normal or sufficiently well behaved that the Tchebyshev inequality holds, then the discussion in all previous chapters concerning the generation of the efficient frontier is useful in finding the optimal portfolio.

MAXIMIZING THE GEOMETRIC MEAN RETURN

One alternative to utility theory is simply to select that portfolio that has the highest expected geometric mean return. Many researchers have put this forth as a universal criterion. That is, they advocate its use without qualifications as to the form of utility function or the characteristics of the probability distribution of security returns. The proponents of the geometric mean usually proceed with one of the following arguments. Consider an investor saving for some purpose in the future, for example, retirement in 20 years. One reasonable portfolio criterion for such an investor would be to select that portfolio that has the highest expected value of terminal wealth. Latane (1959) has shown that this is the portfolio with the highest geometric mean return. The proponents have also argued that the maximum geometric mean[4]

1. has the highest probability of reaching, or exceeding, any given wealth level in the shortest possible time[5]
2. has the highest probability of exceeding any given wealth level over any given period of time[6]

These characteristics of the maximum geometric mean portfolio are extremely appealing and have attracted many advocates. However, maximizing the geometric mean implicitly assumes a particular trade-off between the expected value of wealth and the occurrence of really bad outcomes. It is not clear that maximizing the geometric mean return is always appropriate.

Opponents quickly point out that, in general, maximizing the expected value of terminal wealth (or any of the other benefits discussed earlier) is not identical to maximizing the utility of terminal wealth. Because opponents accept the tenets of utility theory, and, in particular, the idea that investors should maximize the expected utility of terminal wealth, they reject the geometric mean return criteria.

In short, some researchers find the characteristics of the geometric mean return so appealing they accept it as a universal criterion. Others find any criterion that can be inconsistent with expected utility maximization unacceptable. Readers must judge for themselves which of these approaches is more appealing.

Having discussed the arguments in favor of and against the use of the geometric mean as a portfolio selection criterion, let us examine the definition of the geometric mean and some properties of portfolios that maximize the geometric mean criterion. The geometric mean is easy to define. Instead of adding together the observations to obtain the mean, we multiply them. If R_{ij} is the ith possible return on the jth portfolio and each outcome is equally likely, then the geometric mean return on the portfolio $(\overline{R}_{Gj})$ is

$$\overline{R}_{Gj} = (1 + R_{1j})^{1/N}(1 + R_{2j})^{1/N}...(1 + R_{Nj})^{1/N} - 1.0$$

If the likelihood of each observation is different and P_{ij} is the probability of the ith outcome for portfolio j, then the geometric mean return is

$$\overline{R}_{Gj} = (1 + R_{1j})^{P_{1j}}(1 + R_{2j})^{P_{2j}}...(1 + R_{Nj})^{P_{Nj}} - 1.0$$

[4]The accuracy of these statements is not universally accepted.

[5]See Brieman (1960) for a discussion of this property. Roll (1973) argues that this is true only in the limit.

[6]See Brieman (1960) for a discussion of this property. Roll (1973) and Hakansson and Miller (1975) make a similar argument.

Table 11.9 Geometric Mean Returns

	Securities			
Outcome	A	B	C	Portfolio
1	0.80	−0.10	−0.20	0.16 2/3
2	−0.30	0.30	0.60	0.20
Geometric mean	0.12	0.08	0.13	0.18

This is sometimes written in compact form. The symbol π means product. Thus the preceding series can be written as

$$\overline{R}_{Gj} = \prod_{i=1}^{N} (1 + R_{ij})^{P_{ij}} - 1.0$$

The portfolio that has the maximum geometric mean is usually a diversified portfolio. This can be illustrated with an example. Table 11.9 shows three possible investments listed as securities A, B, and C. Each of these investments has two possible outcomes, each equally likely. The portfolio shown consists of equal proportions of each of the three securities. As can be seen from the table, the portfolio has a higher geometric mean return than any of the individual securities. This result is easily explained. The geometric mean return penalizes extreme observations. In fact, a strategy with any probability of bankruptcy would never be selected as it would have a zero geometric mean.[7] As we have seen in other chapters, portfolios have less extreme observations than individual securities. Thus the geometric mean strategy usually leads to a diversified strategy.

While the portfolio that maximizes the geometric mean is likely to be highly diversified, it will not (except in special circumstances) be mean–variance efficient. There are two cases in which mean–variance analysis is meaningful for locating the portfolio with the highest geometric mean return.

First, maximizing the geometric mean return is equivalent to maximizing the expected value of a log utility function.[8] The log utility function is

$$U(w) = ln(w)$$

[7]If one possible outcome is a return of −1, then for that outcome, $(1 + R_{ij}) = (1 - 1) = 0$. The geometric mean is the product of the $(1 + R_{ij})$. The whole product becomes zero if one element is zero. Thus the geometric mean criteria would never select an investment with any probability of bankruptcy.

[8]The log utility function can be written as

$$\max E\ ln(w_1)$$

where w_1 is end of period wealth, a random variable. Because utility functions are unchanged up to a linear transformation, if we let w_0 stand for the funds the investor can invest, then we can write the problem as

$$\begin{aligned} \max E\ [ln(w_1) - ln(w_0)] &= \max E\ ln(w_1 / w_0) \\ &= \max E\ ln(1 + R_1) \\ &= \max \sum_i P_i \times ln(1 + R_i) \\ &= \max \sum_i ln(1 + R_i)^{P_i} \end{aligned}$$

We know from earlier in this chapter that if returns are normally distributed, then mean-variance portfolio analysis is appropriate for investors interested in maximizing expected utility. Investors with log utility functions are such investors. Thus investors interested in maximizing the geometric mean return could use mean–variance analysis if returns were normally distributed.

It has also been shown that the portfolio that maximizes the geometric mean return is mean–variance efficient if returns are log-normally distributed. In this case, a very simple formula exists that indicates which portfolio in the mean-variance efficient set is to be preferred.[9] With the exception of these two cases, the portfolio with the maximum geometric mean return need not be mean–variance efficient.

When returns are not normally, or log-normally, distributed, more general procedures are needed to determine the optimum portfolio. Ziemba (1972) discusses one possible approach. Maier, Peterson, and Vanderweide (1977) discuss a second approach.

VALUE AT RISK (VaR)

Institutions such as banks and insurance companies are concerned with the likelihood of bad outcomes. We have seen one way to express willingness to tolerate bad outcomes in our presentation of safety first. Another widely used approach is value at risk. Safety first involved the trade-off of expected return and a bad outcome. Value at risk looks only at the size of bad outcomes that can occur with a specified probability in a specific time interval. For example, the institution might calculate that there is a 5% probability of a loss of $295,000 or more occurring in the next week. If management were interested in the 5% probability level, then $295,000 would be the value at risk. Let us discuss how this value at risk is determined.

Assume a portfolio is $100 million in value. Assume the expected return over the next week is 0.2%, with a standard deviation of 0.3%. Also assume normal distributions. Then we know that the lowest 5% of possible returns are returns that occur more than 1.65 standard deviations away from the mean. Thus, 5% of the time, we can expect returns below $\bar{R} - 1.65\sigma$ or $0.2 - (1.65)*(0.3)$. Simplifying this results in a return of -0.295% or less. If this investor has 100 million in assets, this is a loss of 295,000 or more. This dollar number, $295,000, is called value at risk (VaR). VaR is the best outcome that can occur if returns are in the worst part of the possible outcomes. If one is willing to assume normal distributions, then all the tools learned in prior chapters are applicable for estimating the mean and standard deviation, and the computation is straightforward. Estimate the mean return and standard deviation over the period in question and use the normal distribution to determine how many standard deviations from the mean you are concerned

---

Because the sum of the logs of a set of variables is the same as the log of the products, this problem can be written as

$$\text{Max} \ln \prod_i (1 + R_i)^{p_i}$$

But this is just the log of 1 plus the geometric mean return. Because taking the log of a set of numbers maintains the rank order, then the portfolio with the highest geometric mean return will also be the preferred portfolio if the investor has a log utility function.

[9]See Elton and Gruber (1981). The optimum portfolio is the one that maximizes

$$\frac{\sigma\bar{R}}{\bar{R}^2 + \sigma^2}$$

with. The worst 5% is the common choice, which, as we discussed earlier, is 1.65 standard deviations from the mean, *t*. This return is then computed (in the example, mean minus 1.65 standard deviations is computed) and multiplied times the value of the assets to get the least dollar loss if returns are in the worst possible set of outcomes (in our example, the lowest 5%). This is how one finds the VaR.

Many institutions hold assets that do not have normal distributions of returns, such as securities with option-like elements. These institutions usually use simulation to compute VaR. Simulation is discussed later in this chapter. These institutions simulate possible return paths thousands of times and then determine the best returns among the bad outcomes. In our example, if the institution performed 1,000 simulations, and they were worried about the worst 5% of outcomes, they would sort the outcomes and, from the 50 worst outcomes (lowest 5%), take the highest return. This, times the assets, results in the dollar loss, and this dollar loss would be designated as the VaR.

UTILITY AND THE EQUITY RISK PREMIUM

Utility theory is a potentially powerful tool for portfolio decision making. All utility functions have a constant that serves to specify the trade-off between risk and return. This constant is called the *coefficient of risk aversion*. Over the past 20 years, the question of utility theory's potential for realistic application has been subject to considerable debate. A major conceptual challenge to utility theory is the equity premium puzzle posed by Mehra and Prescott (1985). The question they ask is a simple one: shouldn't the risk aversion of the average investor imply an equity risk premium approximately equal to its historical value?

Up to this point in the book, we have assumed that the risk and return of assets are given. Later, however, we will introduce the concept of equilibrium models, in which the expected return of an asset is a result of an equilibrium of supply and demand for the asset. In an equilibrium dominated by very risk-averse investors, we should see risky assets like stocks providing a higher expected return to attract cautious investors. In an equilibrium in which the average investor is only mildly risk averse, the spread in expected return between stocks and less risky assets should be much smaller. What you do *not* expect to find is a high expected return relative to less risky assets and mildly risk-averse investors. Yet this is precisely what Mehra and Prescott discovered in their equilibrium analysis of the U.S. capital markets over the period 1889 to 1978. Over that time period, the average annual return to U.S. stocks was 6% per year greater than the return on risk-free debt. Conversely, their theoretical model with a realistic constant relative risk aversion coefficient (between 1 and 2) for the aggregate U.S. investor implies that this equity risk premium should be less than 1% per year. Mehra and Prescott calculated that the coefficient of risk aversion required to generate the historical spread between stock returns and riskless bond returns would have to be between 30 and 40. What does this number mean?

Suppose an individual with a risk aversion of 50 faced a 50-50 gamble of doubling or halving his savings. With this level of risk aversion, he would pay 49% of his savings to avoid the loss of 50%. This kind of behavior is difficult to rationalize. This individual would forgo a 50% chance of doubling his money and accept a certain loss of 49% to avoid losing an additional 1% more.[10]

This discrepancy between the high historical equity premium and the modest one implied by utility theory calls into question either the efficacy of utility models or the validity of historical asset returns. The puzzle has stimulated considerable research into investor utility as well as research into the use of historical data in the assessment of risk

[10]A more complete discussion of the equity premium puzzle can be found in Siegel and Thaler (1997).

and expected return. Attempts to solve the problem can be roughly divided into two classes: empirical and theoretical.

Empirical Solutions

The empirical, data-based approach asks whether the equity premium is properly measured, given the data available to researchers for study. For example, perhaps a century ago, investors believed that stock investing was a very risky prospect, and they demanded a commensurate compensation for holding equities. A century later, their equity investments turned out well—at least as measured by the U.S. historical stock market data. Economist Thomas Rietz (1998) argues that investors in the past may have properly anticipated crashes that just never occurred in the data—but it does not mean that they could not have happened, only that we were lucky enough to avoid such disasters.

In fact, when we look at the returns to the U.S. market, we know we are looking at the lucky market. The United States was on the winning side of the two world wars of the twentieth century and also grew to become the dominant stock market in the world by the late twentieth century. These facts alone would suggest that it is not a representative sample to measure stock market performance or to measure the equity premium. To examine this issue, Brown, Goetzmann, and Ross (1995) develop a simple model in which some stock markets decrease in capitalization and finally disappear, while others thrive and end up in the historical record. They show that by only including those that survive, the estimated historical premium will be positively biased.

Goetzmann and Jorion (1999) collect a database of capital appreciation indexes for 39 markets going back to the 1920s. For 1921 to 1996, U.S. equities had the highest real return of all countries, at 4.3%, versus a median of 0.8% for other countries. The high equity premium obtained for U.S. equities appears to be the exception rather than the rule.

The financial crisis of 2008 and attendant collapse of equity values worldwide suggests that the recent and favorable history of the markets might not be a good basis on which to determine long-term equity premia. Claus and Thomas (2001) and Fama and French (2002) use nonreturn data such as earnings forecasts, dividends, and growth rates to estimate the expected equity risk premium. These approaches may avoid the pitfalls of survival bias.

Theoretical Solutions

Theoretical solutions to the equity premium puzzle, conversely, have led to the development of more sophisticated models of investor utility functions and attitudes toward risk. For example, one class of theoretical solutions posits that rational investors hate to see their standard of living decline, even when it has recently increased. They are thus very averse to even small drops in their wealth. This is called *habit formation* or *ratcheting of consumption*. This way of modeling preferences means that, no matter how much wealth an investor accumulates, she has an extraordinary aversion to a small drop in current wealth level, or in her wealth compared to everyone else's wealth. These models are theoretically sound, but they have some difficulty in explaining the average level of stock market participation—they predict a widespread investment in stocks by people at an early stage of life, a pattern not observed in surveys of household finances.

Another approach seeks to explain the equity premium puzzle as a result of irrational or inconsistent investor choice. Benartzi and Thaler (1995), for example, suggest that the high equity premium may be the result of equity investors focusing too much on short-term market

performance. For example, suppose that you hired a money manager and entrusted this manager with your entire asset portfolio. You also instructed the manager not to lose money in any single year or you would fire him. In this situation, the manager would be unwilling to invest in stocks or would incur large insurance costs to cover the risk of a loss in any given year. The portfolio would simply not grow as quickly as it would have if the manager had been told that a loss in any given year was acceptable, but a loss over a 20-year horizon was not. The effect of evaluating portfolio performance on an annual basis with respect to a given required return is to make investors more averse to equity investment, effectively increasing the premium demanded by investors to hold stocks.

Another approach to explaining the equity premium puzzle is taken by Cogley and Sargent (2008). According to their research, a sufficiently pessimistic prior combined with reasonable Bayesian updates can produce estimates of the equity premium, prices, and returns consistent with those observed in the market.

Although all of these attempts to solve the equity premium puzzle have added significantly to financial research, none has yet satisfactorily reconciled models over investor utility with the empirically observed excess return of stocks over bonds. Until we further understand this divergence between data and theory, it is wise to use utility analysis with some caution. The notion that we may not properly understand even the order of magnitude of aggregate investor risk aversion is troubling.

OPTIMAL INVESTMENT STRATEGIES WITH INVESTOR LIABILITIES

Up to this point, the optimization model has focused primarily on portfolio assets, however practically all portfolios exist to meet some future obligations. Pension funds are set up to provide income and benefits to retirees. Endowments support current and future expenses of universities and foundations. Insurance company portfolios are designed to build assets to meet future claims. In all of these cases, the primary goal of the investor is not simply asset growth but fulfilling future commitments. The investor is thus concerned with the growth of assets net of future outflows. In particular, a financial intermediary may be concerned with changes in net worth, where net is defined in terms of a set of existing liabilities.

There are different ways to express the problem of net worth optimization, however, they all are related to the basic challenge of adapting a potentially complex set of future liabilities to the two-dimensional framework of the portfolio optimization model. In essence, the liabilities faced by the fund must be characterized by expectations of mean return, standard deviation, and correlations of assets if they are to fit into the risk-return space.

Consider, for example, a pension fund that has a known set of cash payouts due in a 5-year period extending from 10 years to 15 years in the future. The efficient frontier technology can be adapted to optimizing the portfolio with respect to these anticipated liabilities. In this case, the riskless asset, from the fund's perspective, would be a portfolio of bonds with cash flows precisely matching the future stream of liabilities. The risk and return and correlations of this matching portfolio of bonds perfectly characterize the liabilities—in this sense, it could be called a "liability asset."

This cash flow–matching portfolio is also said to defease the liabilities. It thus functions much like the riskless asset in the standard model. Once these known liabilities have been defeased, the fund can optimize over the remaining assets.[12] This is equivalent mathematically to treating the liabilities as negative assets (more properly, as

[12]For details of this approach, see Elton and Gruber (1992). For a discussion of the use of this approach in practice, see Leibowitz and Hendrickson (1988).

shorted assets), and constraining the portfolio to hold the "liability asset" in the proportion that the present value of these future liabilities bears to the current value of the assets in the portfolio.

Thus returns on net worth can be expressed in terms of assets and liabilities. If S_t is surplus or net worth (assets minus liabilities), then return on surplus is

$$R_S = \frac{S_{t+1} - S_t}{S_t}$$

S_{t+1} is determined as assets minus liabilities in period $t + 1$. It follows that

$$\left(1 + R_S\right) = \frac{A_{t+1} - L_{t+1}}{S_t} = \frac{A_{t+1}}{S_t} - \frac{L_{t+1}}{S_t}$$

Multiplying the first term by A_t / A_t and the second term by L_t / L_t results in

$$\left(1 + R_S\right) = \frac{A_{t+1}}{A_t}\frac{A_t}{S_t} - \frac{L_{t+1}}{L_t}\frac{L_t}{S_t}$$

$$(1 + R_s) = \left(1 + R_A\right)\frac{A_t}{S_t} - (1 + R_L)\frac{L_t}{S_t}$$

recognizing that $A_t - L_t = S_t$:

$$R_s = R_A \frac{A_t}{S_t} - R_L \frac{L_t}{S_t}$$

One approach to net optimization is to use historical asset returns net of liabilities as an empirical starting point for the analysis. In the previous example, let us assume that the present value of the assets is twice that of the present value of the liabilities. Because a portfolio of intermediate-term zero-coupon government bonds defeases the liabilities, we may estimate the risk, return, and correlations of liability asset R_L using the historical time-series performance of intermediate-term government bonds.[13] We may also estimate the inputs for three asset classes, stocks [S], intermediate-term government bonds [B], and Treasury bills [F], using historical data. Then we transform each return series to the return on net worth by subtracting off the appropriately scaled liability series. Thus our "net" time series, used to calculate inputs to the optimization model, are $R_S - 1/2\ R_L$, $R_B - 1/2\ R_L$, and $R_F - 1/2\ R_L$. The means, standard deviations, and correlations of these three net series are then used to calculate an efficient frontier.[14] What will this frontier look like? Note first that all of the positions of the basic asset classes change as a result of subtracting off the liabilities.

Consider a portfolio entirely invested in B. Because L and B are perfectly correlated to each other, the liabilities are defeased, that is, perfectly hedged through matching cash flows. This only requires half of the assets, however. The remaining half of the assets are then invested in intermediate-term bonds. This asset portfolio now has half the expected return and half the

[13]We may also adjust this historical time series to reflect current expected returns to bonds going forward—as long as we apply these adjustments to the asset inputs as well.

[14]That it is optimum to defease is shown in Elton and Gruber (1992).

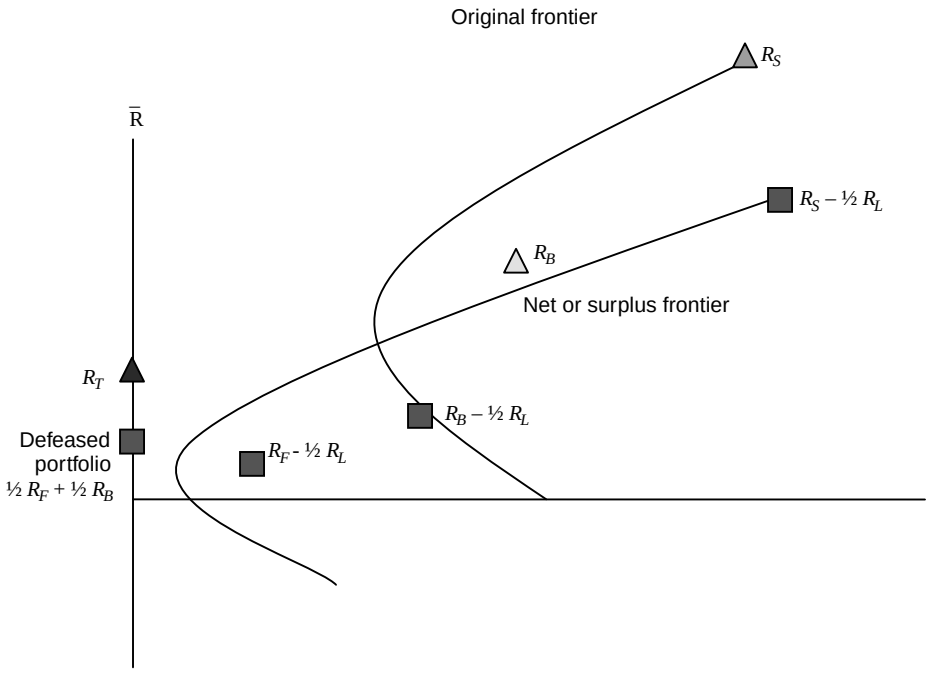


Figure 11.6 Expected return versus variance considering liabilities.

variance compared to what a bond portfolio would have in an "asset-only" optimization because the liabilities have effectively perfectly hedged away this amount of the risk and return from the net investment. This is shown in Figure 11.6. This portfolio, if not held to maturity, has a risk of decreasing in value.

Now consider a portfolio of half T-bills and half bonds. We can express the time series vector of the portfolio P in terms of the vectors of historical net returns,

$$R_P = \tfrac{1}{2}\ [R_F - \tfrac{1}{2}\ R_L] + \tfrac{1}{2}\ [R_B - \tfrac{1}{2}\ R_L]$$

$$R_P = \frac{1}{2} R_F + \frac{1}{2}(R_B - R_L)$$

because $R_B = R_L$:

$$R_P = \frac{1}{2} R_F$$

Thus half the portfolio is placed into bonds that perfectly hedge the liabilities, and the other half is put into the riskless asset. As a result, the portfolio is riskless. This portfolio is shown with an expected return of 1/2 R_F and zero risk in the figure. Notice that a portfolio of 100% T-bills ($R_P = R_F - 1/2\ R_L$) is not riskless. The $R_L/2$ portion has half the risk and minus half the return of the intermediate bond portfolio.

In the preceding example, it is possible to construct a perfect hedge of the portfolio liabilities and to achieve a riskless portfolio by placing the remaining assets in Treasury bills. The problem becomes more challenging when the liabilities are less straightforward to replicate with existing assets and can thus only be approximated with error.

For example, if liabilities are expected to increase with the rate of growth in average wages, there is no existing financial instrument that exactly matches this factor. The manager now must construct a factor from investable assets that is as closely correlated as possible to the liability. In practice, this might be done by regressing the time series of historical growth in average wages on the time series of returns to the financial assets in the investor opportunity set, while constraining the coefficients in the regression to sum to 1.[15]

This achieves an investable portfolio that has two characteristics. First, no other combination of assets better explains the dynamics of the liability—it is the best "hedge" to the liability that can be achieved with a fixed-weight portfolio of assets. Second, the unexplained portion of the liability—the residual risk—is uncorrelated to the assets in the opportunity set. Because it is uncorrelated to the investments, this residual has no influence on the allocation decision.

We can use the preceding example as a starting point to explore this procedure. Let $R_L = (R_B + e)$, where e is an uncorrelated, mean zero error term. Thus R_B cannot perfectly hedge R_L. In this circumstance the all-bond portfolio is

$$R_P = R_B - \frac{(R_L)}{2}$$

Because $R_L = R_B + e$, we have

$$R_P = R_B - \frac{R_B}{2} - \frac{l}{2} = \frac{R_B}{2} - \frac{l}{2}$$

Thus the all-bond portfolio has more risk than in the preceding example. The lower the correlation of B and L, the higher the risk of the 100% B portfolio. Similarly, for the 1/2 T and 1/2 B portfolio,

$$R_P = \frac{1}{2}\left[R_F - \frac{R_L}{2}\right] + \frac{1}{2}\left[R_B - \frac{R_L}{2}\right]$$

$$R = \frac{R_F}{2} + \frac{R_B}{2} - \frac{R_L}{2}$$

and because $R_L = R_B + e$,

$$R_P = \frac{R_F}{2} + \frac{e}{2}$$

Hence the previously riskless portfolio is no longer riskless. An important feature of these portfolios is that no other mix of assets can reduce the additional risk represented by the residual factor $-e/2$. Because it is uncorrelated to the other assets in the portfolio, nothing will provide an additional hedge.

A key feature of the solution to this problem is that, given a situation in which assets are larger than liabilities, the optimizing investor will choose to defease (or match as closely as possible) the entire present value of future liabilities. Failing to do so will leave additional, hedgeable risk. The investor then chooses a portfolio from the remainder of assets according to utility preferences or other means of selecting an optimal portfolio.

[15]The details of this regression can be found in Sharpe (1992).

LIABILITIES AND SAFETY-FIRST PORTFOLIO SELECTION

Once the net optimization approach is chosen, investor preferences must be characterized in terms of risk preferences about the surplus of assets over liabilities. While it may be possible to express these in terms of a utility function, the safety-first approach is useful. For example, it is possible in the safety-first framework to choose a portfolio that minimizes the probability of not meeting the liabilities. The two preceding cases are instructive. When a perfect hedge to liabilities exists, and there are sufficient assets to defease portfolio obligations, there exists a point on the Y axis for which the chance of shortfall is 0%. In the case for which no perfect hedge exists, a point on the efficient frontier can be found that reduces that possibility as much as possible. It is identified by the ray extending from the origin to the point of tangency on the net frontier. This follows from Roy's safety-first criterion. Maximize $R_P - R_L/\sigma_P$, where, in this case, R_L is equal to zero. This is equivalent to the Tangency Portfolio on the efficient frontier with a line passing through zero.

It is possible that the manager of a portfolio with liabilities may not want to maximize the probability of covering liabilities but might instead wish to bear an increased risk of not meeting liabilities in exchange for a higher expected net return. This is exactly a form of the problem solved by Telser. Here the manager selects a probability that he is willing to take that liabilities at the end of the period exceed the assets and maximize expected return for that level. Any probability corresponds to a ray starting at the origin and having a slope determined by the probability level of not meeting liabilities that the manager selects. As explained earlier, this ray is tangent to or crosses the efficient frontier. The portfolio with maximum return that meets the probability goal is obtainable and is defined in Figure 11.7.

SIMULATIONS IN PORTFOLIO CHOICE

One of the limitations of mathematical optimization models is that they do not explicitly allow for changes in portfolio policy or active decision making through time. It is extremely difficult to characterize the whole range of dynamic portfolio strategies one could employ. The range of choices for such strategies is simply too large. Likewise, it is difficult with the optimization model to accommodate intermediate inflows and outflows that depend on past

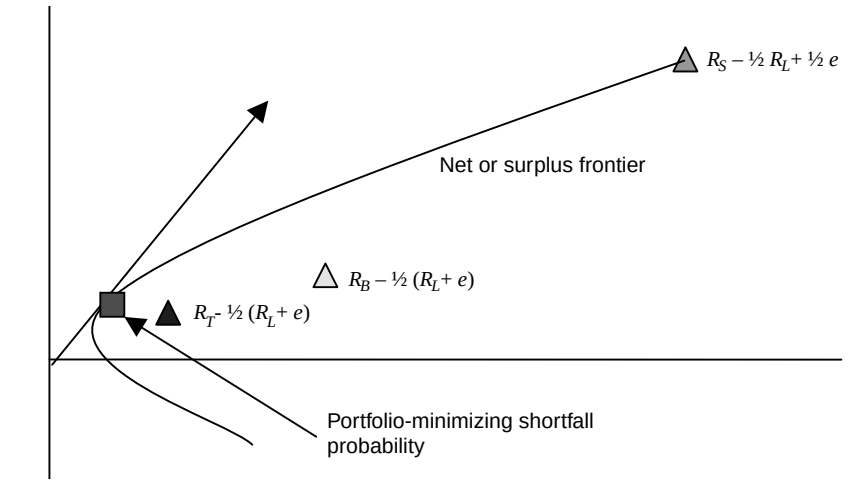


Figure 11.7 Trade-off between maximum return and different probabilities.

events. With dramatic increases in computing power over the past decade, however, it is at least possible to examine the effects of some dynamic strategies and intermediate flows on the performance of the investment portfolio through the use of simulation programs. The concept of simulation is to visualize the range of future outcomes of an investment process by constructing hundreds or even thousands of potential scenarios—each one generated by the same set of investor decision rules and the same set of assumptions about how investment returns behave. In this section, we explore some of these methods.

The first example is an all-stock investment portfolio from which the investor either spends 3% each year or 20% of the profits, whichever is greater. Over the long-term investment horizon, what is the expected distribution of future wealth? To model this, let us assume that U.S. stock market returns from year to year are independent of each other. We can construct a simulated "history"—actually a pseudo-history—of asset returns by randomly drawing from the actual distribution of stock market returns from 1926 to the present with replacement.[16] This particular type of simulation is called a *bootstrap*. Bootstrapping relies on the actual distribution of data rather than on artificial data generated from probability distributions.[17] This is particularly useful when the underlying distribution might be different from the normal or log-normal.

Each single, bootstrapped pseudo-history is now the same length as the original history of the U.S. stock market and has approximately the same risk and return characteristics but an entirely different pattern of growth. Repeating this procedure 50 times generates 50 alternative pseudo-histories that the U.S. market *might* have taken, given its statistical characteristics. Figure 11.8 plots a histogram of the average annual returns to large-cap U.S. stocks for 50 simulations based on these assumptions. The actual average return over this time period was 12.39%, however, the bootstrapped distribution ranges from 7% to 17%, with the most likely outcomes in the 11% to 14% range.

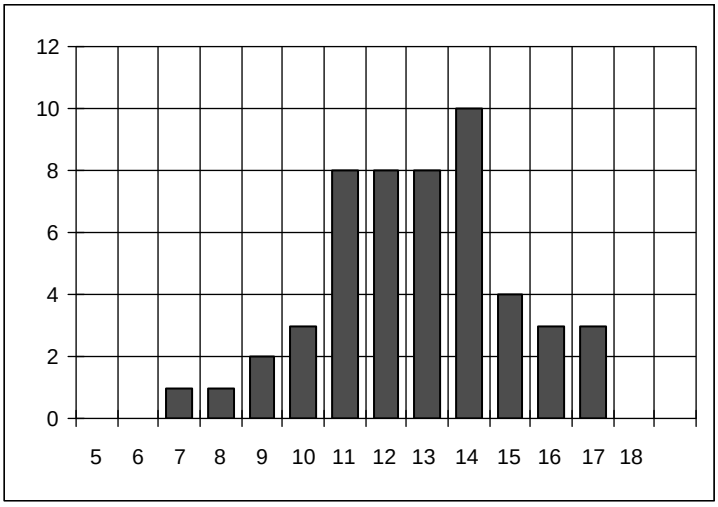


Figure 11.8 Simulated distribution of average returns.

[16]We could also draw from a distribution of returns that matches the actual distribution of returns in terms of mean, standard deviation, skewness, and kurtosis.

[17]See Efron (1979). For applications in finance, see Goetzmann (1990).

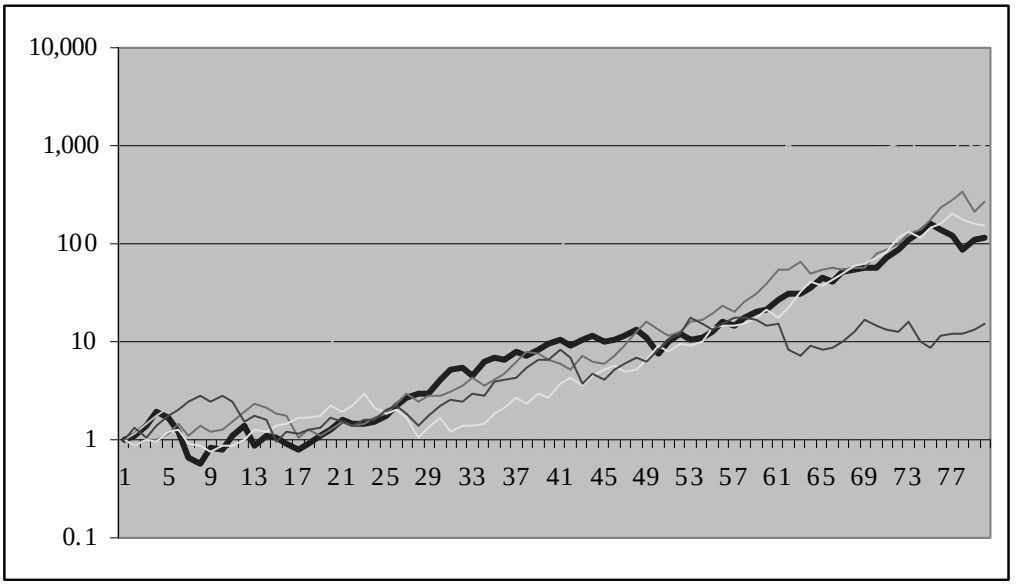


Figure 11.9 Simulated growth of $1 over 79 years.

From each simulated time series of stock market returns, we can calculate the amount an investor would spend each year, depending on the annual return. We may then cumulate these returns net of spending to arrive at a hypothetical 79-year performance record. Each simulation is just one possible outcome that depends on the precise way in which the market behaved in a single simulation.

Figure 11.9 shows the growth of a dollar in wealth over 79 years for 4 of the 50 simulations. The dark line represents the result of applying the spending rule to the actual history of U.S. market returns. The other four lines show what might have happened. They are hypothetical outcomes based on the given assumptions. Like the actual history of U.S. investment performance, these simulated wealth paths have periods of growth and decline. Even though they did not actually happen, they allow us to ask questions of potential interest. For example, how likely is it that the terminal value of the portfolio will exceed $100 in 79 years? We can count the paths that end up above $100 in the figure, or, using all 50 simulated outcomes, we can provide a more accurate probability assessment: 28/50 times or 56%. We can also count other outcomes of interest. For example, how likely is it that the dollar value available for spending under the proposed rule is less than 5% of the portfolio value? How frequently will the portfolio experience a decrease in value over any given five-year interval? These and other questions can be addressed through the simple process of counting up the number of such occurrences in the simulations.

Multiple-Asset Bootstrapping

The bootstrap can be easily applied to multiple assets. Suppose we wished to simulate a 50-50 portfolio of U.S. stocks and non-U.S. equities over time. An important feature of this problem is preserving the correlation structure between the two assets. As long as we can safely assume that the correlation between the two is stable through time, we can perform a bootstrap that simultaneously draws *pairs* of returns together. The way to perform

this multiple-asset bootstrap is to first record the return on U.S. stocks over each year of the 79-year history and the corresponding return for non-U.S. stocks. Then 79 random numbers are drawn from 1 through 79, where each number designates a pair of returns and any number can occur multiple times (random selection with replacement). The 79 matched pairs of returns that correspond to the 79 random numbers is a single pseudo-history of U.S. and non-U.S. stock market performance that preserves not only the statistical characteristics of each individual series but also the correlation between the two. A time series for the 50-50 portfolio return can be computed by averaging the returns of the two assets for each year.

Biased Bootstrapping and Scenario Analysis

The implicit assumption of the basic bootstrapping model is that each actual year's performance (or time period) is equally likely to occur in the pseudo-history. For example, the U.S. stock market return for 1929 has a 1/79 chance of occurring in any given year of a bootstrapped history—indeed, it could occur twice, three times, or not at all, depending on the luck of the draw. Sometimes it is useful for the purposes of analysis to relax the assumption of equally likely draws. For example, suppose we believed the U.S. economy was headed into a period of high inflation, and we wished to examine the expected distribution of the stock market in a high-inflationary environment. One approach to this is to limit the years from which the bootstrap is drawn to high inflation years.

A less drastic adjustment could be made to conform to assumptions about expected future probability of entering a high-inflation economic environment. Instead of allowing returns in years of high inflation and low inflation to have the same probability of being drawn, you can adjust these probabilities by changing the number of high- versus low-inflation years in the population from which you sample. To make high-inflation years twice as likely as low-inflation years, simply include the high-inflation years twice in the population. This will cause them to be sampled twice as frequently in each bootstrapped time series.

Time Series Dependence

The most important assumption of the bootstrap is the independence of returns through time. Unfortunately, there are many financial times series for which this is not a reasonable assumption. Treasury bill returns, for example, have a high degree of autocorrelation from one year to the next (the value next year is highly dependent on this year's value). Inflation has this characteristic as well. It is possible to address this problem, however, it requires some additional statistical modeling and estimation.

Suppose, for example, we wished to create a pseudo-history of U.S. inflation. First, we must estimate a model of the inflation process I:

$$I_t = a + bI_{t-1} + e_t$$

where I_{t-1} is last year's inflation rate and t ranges from 1926 to 2007. For example, if t = 1928, then I_{t-1} is inflation in 1927. The coefficients a and b can be estimated from a linear regression, and e_t is the error term from the regression for year t. Let e_t* be an error term drawn randomly with replacement from the actual residuals for 1926 to 2007 (79 residuals).[18] Let I^*_t be the variable we use to indicate the bootstrapped value of inflation at time t. To construct the bootstrapped series, we begin with an actual starting value, I_{1926}, the inflation

[18]The assumption made here is that e_t is independent of the level of inflation I_t. See Ibbotson and Singuefield (1976).

rate in 1926. I^*_{1927} is calculated as $a + bI_{1926} + e_t^*$, where e_t^* is drawn from the 79 residuals. The next bootstrapped inflation year in the sequence builds on the previous value: $I^*_{1928} = a + b\, I^*_{1927} + e_{t+1}^*$, where this e_{t+1}^* is drawn with replacement for the period $t + 1$ from the 76 regression residuals as before. This process continues until an entire 79-year pseudo-history of inflation is constructed. This method, based on bootstrapping the errors in the autocorrelation model, now preserves the time series dependency of annual inflation, as well as its approximate historical mean and standard deviation. The methodology can also be easily combined with a multiple-asset bootstrap to preserve the correlation between asset returns and autocorrelated series such as inflation or Treasury bills.

Bootstrapping Applications

How well does the bootstrap perform? An early application of bootstrapping and simulation to investment analysis allows us to compare the forecasted returns to their actual realizations. Ibbotson and Sinquefield (1976) used the 50 years of U.S. capital market returns from 1926 through 1974 to estimate the distribution of long-term returns to U.S. stocks, bonds, Treasury bills, and inflation over the period 1976 to 2000 (Table 11.10). They used a procedure similar to that described previously to control for time series dependencies, and they also made some additional assumptions about the effect of the yield curve on expected returns to stocks and bonds.

Because the returns for each asset class over the 1976–2000 period are now known, we can compare actual values to estimated distributions. For example, the median forecast geometric return for stocks was 13.1%, compared to an actual value of 15.3%. This actual value corresponds to the seventieth percentile in the bootstrapped distribution—not unusual in the forecast distribution. The forecast for the riskless rate was 6.9%—also close to the median of the bootstrapped distribution. Inflation was overestimated in the simulations, and bond returns were underestimated. On balance, the simulation performed remarkably well and provides a convincing argument for applying it to the problem of understanding asset return distributions going forward over the next 25 years.

Applications

Value at Risk Notice that Table 11.10 reports extreme percentiles. These percentiles can be useful in estimating VaR because they help the analyst understand the probability of a loss of a given magnitude at a given time horizon. Simulations of this form are part of the tool kit used by investment analysts to calculate VaR for portfolios.

Dynamic choice Bootstrapping and simulations are useful to managers seeking to better understand the effects of decisions such as the choice of a spending rule or the influence on asset allocation decisions of future investment outcomes. It is particularly useful in settings for which a specific mathematical formulation of the problem is difficult.

Taxes An important application is the taxable investment portfolio. Dynamic strategies such as selling losing stocks and not recognizing capital gains can help investors minimize tax liabilities. These rules can be modeled through the use of simulation tools, but they are difficult to include in a portfolio optimization program. Thus a simulation approach may help in the tax-planning process.

In all of these applications, it is important to stress that simulation is not an optimization process per se, in that it does not explicitly rank choices according to a single utility criterion. However, it is a critical tool for investment planning and provides potentially detailed and accurate answers to questions about future return distributions and future investment policies.

Table 11.10 Simulated Total Return Distributions for the Period 1976–2000: Geometric Average Annual Rates (in %), Selected Percentiles, All Series

Percentile*	$\tilde{R}_m$	$\tilde{R}_g$	$\tilde{R}_e$	$\tilde{R}_f$	$\tilde{R}_I$	$\tilde{R}_p$	$\tilde{R}_L$	$\tilde{R}_d$	$\tilde{R}_r$	$\tilde{R}_{mr}$	$\tilde{R}_{gr}$	$\tilde{R}_{or}$
1	2.1	3.1	3.3	2.1	−0.4	−4.6	−1.1	−1.3	−4.8	−5.4	−2.2	−2.5
2	4.3	3.9	4.0	3.3	1.7	−2.8	−0.9	−1.1	−4.1	−2.2	−1.7	−1.7
3	4.9	4.2	4.3	3.4	1.8	−2.1	−0.7	−0.9	−3.8	−1.8	−1.4	−1.6
4	4.3	4.5	3.5	2.0	1.8	−0.6	−0.8	−3.4	−1.3	−1.2	−1.5	−1.6
5	4.8	5.2	4.1	2.0	0.2	−0.3	−0.6	−2.5	−0.1	−0.6	−0.4	−1.5
16	4.8	5.2	4.1	2.9	0.2	−0.3	−0.6	−2.5	0.1	−0.6	−0.4	−0.4
20	6.1	6.1	4.9	4.0	2.5	0.1	−0.3	−1.3	2.7	0.1	0.4	0.4
30	10.3	6.6	7.0	5.6	4.7	3.8	0.5	−0.1	−0.7	4.3	0.7	0.9
40	11.8	7.3	7.5	6.2	5.4	4.9	0.9	0.1	−0.2	5.0	1.1	1.4
50	13.1	7.8	8.1	6.7	6.1	5.8	1.1	0.2	0.3	6.2	1.6	1.8
60	14.0	8.0	8.7	7.3	6.9	7.2	1.4	0.5	0.8	7.3	2.0	2.3
70	15.2	9.1	9.4	7.8	7.8	8.6	1.7	0.6	1.2	8.6	2.5	2.8
80	16.9	9.9	10.2	8.6	8.8	10.0	2.0	0.8	1.6	10.0	3.0	3.4
90	19.3	11.0	11.4	9.6	10.2	11.6	2.5	1.2	2.5	12.7	3.5	4.0
95	21.5	12.3	12.7	10.7	11.3	13.6	2.8	1.4	3.1	14.0	4.0	4.4
96	22.1	12.6	13.0	10.9	11.5	13.8	3.0	1.5	3.2	14.5	4.3	4.6
97	23.5	13.0	13.4	11.0	11.8	14.4	3.2	1.5	3.4	15.2	4.2	4.9
98	24.6	13.1	13.9	11.2	12.5	15.5	3.3	1.7	3.5	15.9	4.5	5.2
99	25.0	13.7	14.3	12.1	13.6	16.3	3.5	1.9	4.0	16.9	5.1	5.7
Mean	13.0	8.0	8.2	6.8	6.4	6.0	1.1	0.3	0.1	6.3	1.5	1.8
S.D	4.9	2.3	2.4	2.1	2.9	4.5	1.1	0.7	1.9	4.6	1.6	1.7

*Each series is simulated for $k = 1{,}399$ for each year over the period. The qth percentile lists the $k/4$the sorted simulation. Even though many of the simulated series are interrelated, each series is sorted independently of the others. Thus the simulated distribution of any one series is conditional on an observation or the distribution of any other series is not ascertainable from this table.

CONCLUSION

In this chapter we have analyzed a number of techniques for selecting the optimum portfolio. Which of these techniques is preferred is a choice the investor has to make. The investor can be comforted by knowing that portfolios that lie near to each other on the efficient frontier are similar in their characteristics. Thus an investor somewhat unsure of exactly which portfolio is best still has a good chance of selecting a portfolio near optimum.

APPENDIX

THE ECONOMIC PROPERTIES OF UTILITY FUNCTIONS

The first restriction placed on a utility function is that it be consistent with more being preferred to less. This attribute, known in the economic literature as *nonsatiation*, simply says that the utility of more ($X + 1$) dollars is always higher than the utility of less (X) dollars. Thus, if we want to choose between two certain investments, we always take the one with the largest outcome. If we are concerned with end-of-period wealth, this property states that more wealth is always preferred to less wealth. If utility increases as wealth increases, then the first derivative of utility, with respect to wealth, is positive. Thus the first restriction placed on the utility function is a positive first derivative.

The second property of a utility function is an assumption about an investor's taste for risk. Three assumptions are possible: the investor is averse to risk, the investor is neutral toward risk, or the investor seeks risk. Risk aversion, risk neutrality, and risk seeking can all be defined in terms of a fair gamble. Consider the gambles (options) shown in Table 11.11.

The option "invest" has an expected value of $(1/2)(2) + (1/2)(0) = \$1$. Assume that an investor would have to pay \$1 to undertake this investment and obtain these outcomes. Thus, if the investor chooses not to invest, the \$1 is kept. This is the alternative: do not invest. The expected value of the gamble is exactly equal to the cost. The position of the investor may be improved or hurt by undertaking the investment, but the expectation is that there will be no change in position. Because the expected value of the gamble shown in Table 11.11 is equal to its cost, it is called a *fair gamble*.

Risk aversion means that an investor will reject a fair gamble. In terms of Table 11.11, it means \$1 for certain will be preferred to an equal chance of \$2 or \$0. Risk aversion implies that the second derivative of utility, with respect to wealth, is negative. If $U(W)$ is the utility function and $U''(W)$ is the second derivative, then risk aversion is usually equated with an assumption that $U''(W) < 0$. Let us examine why this is true.

If an investor prefers not to invest, then the expected utility of not investing must be higher than the expected utility of investing, or

$$U(1) > \frac{1}{2}\ U(2) + \frac{1}{2}\ U(0)$$

Multiplying both sides by 2 and rearranging, we have

$$U(1) - U(0) > U(2) - U(1)$$

Table 11.11 An Example of a Fair Gamble

Invest		Do Not Invest	
Outcome	Probability	Outcome	Probability
2	1/2	1	1
0	1/2		

Examine the preceding expression. The expression means that a one-unit change from 0 to 1 is more valuable than a one-unit change from 1 to 2. This latter change involves larger values of outcomes. A function where an additional unit increase is less valuable than the last unit increase is a function with a negative second derivative.

The assumption of risk aversion means an investor will reject a fair gamble because the disutility of the loss is greater than the utility of an equivalent gain. Functions that exhibit this property must have a negative second derivative. Therefore the rejection of a fair gamble implies a negative second derivative.

Risk neutrality means that an investor is indifferent to whether a fair gamble is undertaken. In the context of Table 11.11, a risk-neutral investor would be indifferent to whether an investment was made. Risk neutrality implies a zero second derivative.

Figures 11.10a and 11.10b show preference functions exhibiting alternative properties with respect to risk aversion. Figure 11.10a presents the shape of utility functions in utility of wealth space that exhibit risk aversion, risk neutrality, and risk preference. Figure 11.10b presents the shape of the indifference curves in expected return standard deviation space

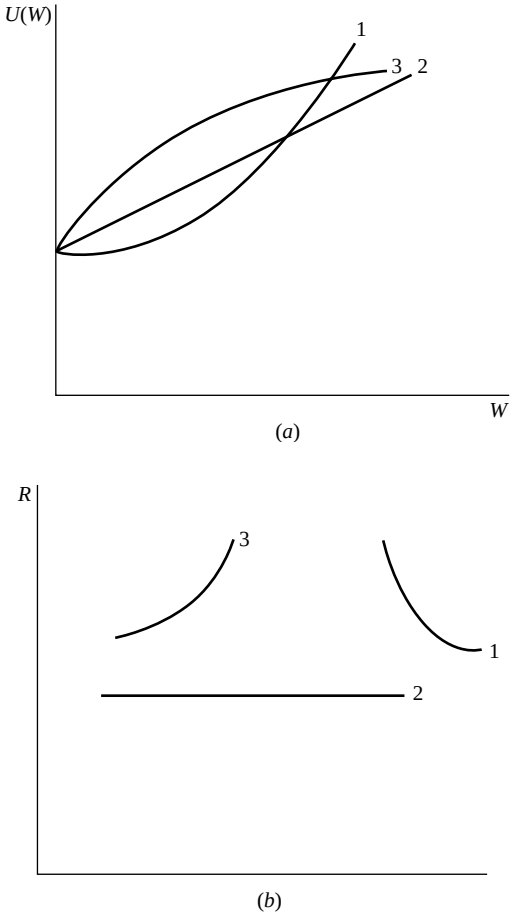


Figure 11.10 Characteristics of functions with different risk-aversion coefficients. (1) Utility function of a risk-seeking investor. (2) Utility function of a risk-neutral investor. (3) Utility function of a risk-averse investor.

Table 11.12 Implications of Attitude toward Risk

Condition	Definition	Implication
1. Risk aversion	Reject fair gamble	$U''(0) < 0$
2. Risk neutrality	Indifferent to fair gamble	$U''(0) = 0$
3. Risk preference	Select a fair gamble	$U''(0) > 0$

that would be associated with each of these three types of utility functions. Table 11.12 summarizes the relationship between risk attitudes and utility functions.

RELATIVE RISK AVERSION AND WEALTH

The third property of utility functions is an assumption about how the investor's preferences change with a change in wealth. Usually the issue is expressed as what percentage is invested in risky assets as wealth changes. For example, if the investor puts 50% of her wealth in risky investments when her wealth is $10,000, does she still put 50% of her wealth in risky assets when her wealth increases to $20,000? If she does, then the investor's behavior is said to be characterized by constant relative risk aversion. If she invests a greater percentage of her wealth in risky investments, she is said to exhibit decreasing relative risk aversion, and if she invests a smaller percentage, she is said to exhibit increasing relative risk aversion. The function $U(w) = \ln(w)$ is frequently used in finance because it exhibits constant relative risk aversion.

If we are able to specify our feeling toward preferring more to less risk aversion and relative risk aversion, we can severely restrict the form of the utility function to consider and will have placed some restrictions on the value the parameters of the utility function can take on. However, there are still many choices, and most investors have difficulty making these choices.

QUESTIONS AND PROBLEMS

1. Consider the following three investments. Which are preferred if $U(W) = W - (1/2)W^2$?

Investment A		Investment B		Investment C	
$ Outcome	Probability	$ Outcome	Probability	$ Outcome	Probability
5	1/3	4	1/4	1	1/5
6	1/3	7	1/2	9	3/5
9	1/3	10	1/4	18	1/5

2. Assume the utility function is $U(W) = -W^{-1/2}$. What is the preferred investment in Problem 1?

3. Consider the following two investments. Which is preferred if the utility function is $U(W) = -W - 0.04W^2$?

Investment A		Investment B	
$ Outcome	Probability	$ Outcome	Probability
7	2/5	5	1/2
10	1/5	12	1/4
14	2/5	20	1/4

4. Consider the choice shown in Problem 3. The probability of a \$5 return is 1/2 and of a \$12 return is 1/4. How much would these probabilities have to change so that the investor is indifferent between investments A and B?

5. Consider the following investments. Which is preferred if $U(\mathrm{W}) = W - 0.05W^2$?

Investment A		Investment B	
Outcome	Probability	Outcome	Probability
5	0.20	6	0.30
7	0.50	8	0.60
10	0.30	9	0.10

6. In Problem 5, what is the minimum amount that the \$5 outcome would have to be changed to so that the investor is indifferent between the two investments?

7. If R_L equals 5%, what is the preferred investment shown in Problem 1 using Roy's safety-first criterion?

8. If α equals 10%, what is the preferred investment shown in Problem 1 using Kataoka's safety-first criterion?

9. If $R_L = 5\%$ and $\alpha = 10\%$, what is the preferred investment shown in Problem 1 using Telser's safety-first criterion?

10. Using geometric mean return as a criterion, which investment is to be preferred in Problem 1?

11. Given the following investments, if R_L is 3%, what investment is preferred using Roy's safety-first criterion?

A		B		C	
Probability	Outcome %	Probability	Outcome %	Probability	Outcome %
0.4	3	0.1	5	0.1	5
0.3	4	0.2	6	0.1	7
0.1	6	0.1	8	0.2	8
0.1	7	0.2	9	0.2	9
0.1	9	0.4	10	0.4	11

12. Using geometric mean, which investment is preferred in Problem 11?

BIBLIOGRAPHY

1. Artzner, P., Delbaen, F., Eber, J. M., and Heath, D. "Coherent Measures of Risk," *Mathematical Finance*, **9,** No. 2 (1999), pp. 203–228.
2. Bakshi, Gurdip S., and Chen, Zhiwu. "The Spirit of Capitalism and Stock-Market Prices," *American Economic Review*, **86,** No. 1 (1996), pp. 1233–1257.
3. Benartzi, Shlomo, and Thaler, Richard H. "Myopic Loss Aversion and the Equity Premium Puzzle," *Quarterly Journal of Economics*, **110,** No. 1 (1995), pp. 73–92.
4. Bernoulli, Daniel. "Exposition of a New Theory on the Measurement of Risk," *Econometrica*, **32,** No. 1 (1954), pp. 23–26.
5. Brennan, Michael J., and Schwartz, Edward S. "On the Geometric Mean Index: A Note," *Journal of Financial and Quantitative Analysis*, **XX,** No. 1 (March 1985), pp. 119–122.

6. Brieman, Leon. "Investment Policies for Expanding Businesses Optimal in a Long Run Sense," *Naval Research Logistics Quarterly*, **7** (Dec. 1960), pp. 647–651.
7. Brown, Stephen J., Goetzmann, William, and Ross, Stephen. "Survival," *Journal of Finance*, **50,** No. 3 (1995), pp. 853–874.
8. Campbell, John Y. "Consumption and Portfolio Decisions When Expected Returns Are Time Varying," *Quarterly Journal of Economics*, **114,** No. 2 (1999), pp. 433–492.
9. Campbell, John Y., and Cochrane, John H. "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, **107,** No. 1 (1995), pp. 205–251.
10. Campbell, Rachel. "Optimal Portfolio Selection in a Value-at-Risk Framework," *Journal of Banking and Finance*, **25,** No. 9 (Sept. 2001), pp. 1789–1804.
11. Claus, James, and Thomas, Jacob. "Equity Premia as Low as Three Percent? Evidence from Analysts' Earnings Forecasts for Domestic and International Stock Markets," *Journal of Finance*, **56,** No. 5 (2001), pp. 1629–1666.
12. Cogley, Timothy, and Sargent, Thomas J. "The Market Price of Risk and the Equity Premium: A Legacy of the Great Depression," *Journal of Monetary Economics*, **55,** No. 3 (2008), pp. 454–476.
13. Constantinides, George M. "Habit Formation: A Resolution of the Equity Premium Puzzle," *Journal of Political Economy*, **98,** No. 3 (1990), pp. 519–543.
14. Dybvig, Philip H. "Dusenberry's Ratcheting of Consumption: Optimal Dynamic Consumption and Investment Given Intolerance for Any Decline in Standard of Living," *Review of Economic Studies*, **62,** No. 2 (1995), pp. 287–313.
15. Efron, B. "Bootstrap Methods: Another Look at the Jacknife," *Annals of Statistics*, **7,** No. 1 (1979), pp. 1–26.
16. Ekern, Steinar. "Time Dominance Efficiency Analysis," *Journal of Finance*, **36,** No. 5 (Dec. 1981), pp. 1023–1034.
17. Elton, Edwin J., and Gruber, Martin J. "Optimal Investment Strategies with Investor Liabilities," *Journal of Banking and Finance*, **16,** No. 4 (1992), pp. 869–890.
18. Elton, Edwin J., and Gruber, Martin J. "On the Optimality of Some Multiperiod Portfolio Selection Criteria," *Journal of Business*, **47,** No. 2 (April 1974), pp. 231–243.
19. Elton, Edwin J., and Gruber, Martin J. "An Algorithm for Maximizing the Geometric Mean," *Management Science*, **21,** No. 4 (Dec. 1974), pp. 483–488.
20. Fama, Eugene, and French, Kenneth R. "The Equity Premium," *Journal of Finance*, **57,** No. 2 (2002), pp. 637–659.
21. Friedman, Milton, and Savage, Leonard J. "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, **48,** No. 2 (1948), pp. 279–304.
22. Goetzmann, William N. "Bootstrapping and Simulation Tests of Long-Term Patterns in Stock Market Behavior," Ph.D. Thesis, Yale University (1990).
23. Gomes, Francisco J., and Michaelides, Alexander. "Portfolio Choice with Internal Habit Formation: A Life-Cycle Model with Uninsurable Labour Income Risk," *CEPR Discussion Papers* 3868, E.E.P.R. (2003).
24. Green, Richard C. "Positively Weighted Portfolios on the Minimum-Variance Frontier," *Journal of Finance*, **41,** No. 5 (Dec. 1986), pp. 1051–1068.
25. Hakansson, Nils. "Capital Growth and the Mean-Variance Approach to Portfolio Selection," *Journal of Financial and Quantitative Analysis*, **VI,** No. 1 (Jan. 1971), pp. 517–557.
26. ——. "Comment on Merton and Samuelson," *Journal of Financial Economics*, **1,** No. 1 (May 1974), pp. 950–970.
27. Hakansson, Nils, and Ching, Liu Tien. "Optimal Growth Portfolios When Yields Are Serially Correlated," *Review of Economics and Statistics*, **LII,** No. 4 (Nov. 1970), pp. 385–394.
28. Hakansson, Nils, and Miller, Bruce. "Compound-Return Mean–Variance Portfolios Never Risk Ruin," *Management Science*, **22,** No. 4 (Dec. 1975), pp. 391–400.
29. Ibbotson, Roger G., and Sinquefield, Rex A. "Stocks, Bonds, Bills, and Inflation: Simulations of the Future (1976–2000)," *Journal of Business*, **49,** No. 3 (1976), pp. 313–338.

30. Jean, William H., and Helms, Billy P. "Geometric Mean Approximations," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 3 (Sept. 1983), pp. 287–294.
31. Jorion, Phillipe, and Goetzmann, William N. "Global Stock Markets in the Twentieth Century," *Journal of Finance*, **54,** No. 3 (1999), pp. 953–980.
32. Kritzmann, Mark, and Rich, Don. "Beware of Dogma," *Journal of Portfolio Management*, **24,** No. 4 (Summer 1998), pp. 66–67.
33. Kroll, Yoram, Levi, Haim, and Markovits, Henry M. "Mean–Variance versus Direct Utility Maximization," *Journal of Finance*, **39,** No. 1 (1984), pp. 47–60.
34. Latane, Henry. "Criteria for Choice among Risky Ventures," *Journal of Political Economy*, **59,** No. 1 (April 1959), pp. 144–155.
35. Latane, Henry, and Young, Williams E. "Test of Portfolio Building Rules," *Journal of Finance*, **XXIV,** No. 4 (Sept. 1969), pp. 595–612.
36. Leibowitz, Martin J., and Hendrickson, R. D. "Portfolio Optimization within a Surplus Framework," *Financial Analyst's Journal*, **44,** No. 2 (1988), pp. 43–51.
37. Litzenberger, Robert, and Budd, A. P. "A Note on Geometric Mean Portfolio Selection and the Market Prices of Equities," *Journal of Financial and Quantitative Analysis*, **VI,** No. 5 (Dec. 1971), pp. 1277–1282.
38. Maier, Steven, Peterson, David, and Vanderweide, James. "A Monte Carlo Investigation of Characteristics of Optimal Geometric Mean Portfolios," *Journal of Financial and Quantitative Analysis*, **XII,** No. 2 (June 1977), pp. 215–233.
39. Markowitz, Harry. *Portfolio Selection Efficient Diversification of Investments* (New York: John Wiley, 1959).
40. Mehra, Rajneesh, and Prescott, Edward C. "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, **15 (**1985), pp. 145–161.
41. Mossin, Jan. "Investment for the Long-Run: New Evidence for an Old Rule," *Journal of Finance*, **XXXI,** No. 5 (1976), pp. 1273–1286.
42. ——. *Theory of Financial Markets* (Englewood Cliffs, NJ: Prentice Hall, 1973).
43. Ohlson, James. "Quadratic Approximations of the Portfolio Selection Problem When the Means and Variances Are Infinite," *Management Science*, **23,** No. 6 (Feb. 1977), pp. 576–584.
44. Pratt, J. "Risk Aversion in the Small and in the Large," *Econometrica*, **42,** No. 1 (1964), pp. 122–136.
45. Price, Kelly, Price, Barbara, and Nantell, Timothy J. "Variance and Lower Partial Moment Measures of Systematic Risk: Some Analytical and Empirical Results," *Journal of Finance*, **37,** No. 3 (June 1982), pp. 843–906.
46. Pyle, David, and Turnovsky, Stephen. "Safety-First and Expected Utility Maximization in Mean-Standard Deviation Portfolio Analysis," *Review of Economics and Statistics*, **LII,** No. 1 (Feb. 1970), pp. 75–81.
47. Rietz, Thomas A. "The Equity Premium: A Solution," *Journal of Monetary Economics*, **15,** No. 1 (1998), pp. 145–162.
48. Roll, Richard. "Evidence on the 'Growth-Optimum' Model," *Journal of Finance*, **XXVIII,** No. 3 (June 1973), pp. 551–556.
49. Ross, Stephen A. "Adding Risks: Samuelson's Fallacy of Large Numbers Revisited," *Journal of Financial and Quantitative Analysis*, **34,** No. 3 (Sept. 1999), pp. 323–339.
50. Roy, A. D. "Safety-First and the Holding of Assets," *Econometrics*, **20** (July 1952), pp. 431–449.
51. Samuelson, Paul. "The Fundamental Approximation Theorem of Portfolio Analysis in Terms of Means Variances and Higher Moments," *Review of Economic Studies*, **25,** No. 1 (Feb. 1958), pp. 65–86.
52. Sharpe, William F. "Asset Allocation: Management Style and Performance Measurement," *Journal of Portfolio Management*, **30,** No. 10 (1992), pp. 7–16.
53. Siegel, Jeremy J., and Thaler, Richard H. "Anomalies: The Equity Premium Puzzle," *Journal of Economic Perspectives*, **11,** No. 1 (1997), pp. 191–200.

54. Tehranian, Hassan, and Helms, Billy P. "An Empirical Comparison of Stochastic Dominance among Lognormal Prospects," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 2 (June 1982), pp. 217–226.
55. Vanderweide, James, Peterson, David, and Maier, Steven. "A Strategy Which Maximizes the Geometric Mean Return on Portfolio Investments," *Management Science*, **23,** No. 10 (June 1977), pp. 1117–1123.
56. ——. "Reply to Aucamp," *Management Science*, **24,** No. 8 (April 1978), p. 859.
57. Von Neumann, J., and Morgenstern, O. *Theory of Games and Economic Behavior*, 2nd ed. (Princeton, NJ: Princeton University Press, 1947).
58. Young, Williams, and Trent, Roberts. "Geometric Mean Approximations of Individual Securities and Portfolio Performance," *Journal of Financial and Quantitative Analysis*, **IV,** No. 2 (June 1969), pp. 179–199.
59. Ziemba, William. "Note on 'Optimal Growth Portfolios When Yields Are Serially Correlated,'" *Journal of Financial and Quantitative Analysis*, **VII,** No. 4 (Sept. 1972), pp. 1995–2000.

Section 4

Widening the Selection Universe

12

International Diversification

The investment models developed in the preceding chapter are often applied to and tested on U.S. capital market data. In fact, investors face a much broader opportunity set. International investing has a very long history, particularly in Europe, where foreign participation in the fixed income and equity markets has been active for nearly three centuries. By contrast, for much of the last century, American investors, as well as those of several other countries, manifested a well-documented "home country bias," which was seen an empirical puzzle. Proposed solutions to the puzzle included barriers to cross-border investing as well as behavioral biases.[1]

With increasing globalization of capital markets over the last several decades, U.S. and Asian investors have gained more access to international markets, and this has made the question of how best to diversify all the more important. In this chapter we discuss cross-border investing from the perspective of market integration, the benefits of diversification, and the exposure to institutional risks and frictions. We present ways to calculate the expected returns to cross-border investments, the effect of exchange rates, and approaches to hedging foreign exchange risk.

By almost any measure, cross-border investing has grown dramatically in the last two decades. A 2011 McKinsey report on world capital markets observes that "investment in foreign assets [in 2010] reached $96 trillion, nearly ten times the amount in 1990."[2] The same report estimated the value of world equity at $54 trillion, roughly a quarter of the world's total financial assets of $212 trillion in equity and debt and loans. As of 2010, roughly 32% of the $212 trillion were U.S. assets, 39% were Western European and other developed nations, 12% Japanese, 8% Chinese, and the remaining 9% emerging market assets. Of these, emerging markets have been the fastest growing subset of securities. These dramatic figures show that the opportunity set for any investor in the world is dominated by cross-border choices, and that these choices are changing with the different rates of growth in countries around the world.

[1]Black (1974), French and Poterba (1990), and Huberman (2001).
[2]Roxbugh Lund and Piotrowski (2011).

HISTORICAL BACKGROUND

International cross-border investment dates back to 1720 and the first global stock market bubble. In that year, Dutch, French, and British investors speculated in stocks in London, Paris, and Amsterdam on expectations about profitable trade in the New World. The ensuing crash dropped share prices by 90% and spread from one city to another, in part due to the capital flows of foreign investors. The European markets survived the crisis, however, and eventually became the source of much of the world's financial capital over the next two centuries.

For example, investors in the nineteenth century in London, Amsterdam, Paris, Berlin, and Brussels financed South American railroads, Russian oil companies, the Suez Canal, Chinese banks, African mines, and a host of other nondomestic firms. The investment portfolio of the average London-based investor at that time placed heavy emphasis on investment in non-British assets.[3] One of first texts on portfolio theory, called "Investment an Exact Science," was written in 1907 as a guide to instruct British investors in the early twentieth century how to take advantage of global diversification.[4] Its author, Henry Lowenfeld, noted that the stocks in one country tended to move together, however, stocks in different countries did not. This, he argued, allowed a cross-border investor to reduce the risk of his portfolio. Charts from his book (Figures 12.1a and b) are instructive. He first used graphs to discover that stocks within each country moved together, while stocks in different countries moved differently. He then showed that diversifying across geographical regions, even within one industry, was less risky than diversifying across industries within a single country.

Cross-border investing declined during the middle of the twentieth century with the effects of World War II and geopolitical barriers to foreign investing. Many of the countries Lowenfeld studied are no longer accessible to foreign investors after 1945. Only late in the twentieth century did international capital markets begin to revive. With the fall of the Berlin wall in 1990, the growth of capital markets in many formally Socialist countries was possible. With this revival, individual investors began to take a strong interest in nondomestic investment. Early empirical studies of international investing showed that portfolio risk could be dramatically reduced by allocation across several countries.[5] The large risk reductions potentially achievable led to the question of why every investor did not take full advantage of the opportunity to invest cross-border. To examine this question, we need to analyze the correlation between markets and the risk and return of each market. But before we do this, we must first examine how to calculate returns on foreign investments.

CALCULATING THE RETURN ON FOREIGN INVESTMENTS

The return on a foreign investment is affected by the return on the assets within its own market and the change in the exchange rate between the security's own currency and the currency of the purchaser's home country. Thus the return on a foreign investment can be quite different than simply the return in the asset's own market and can differ according to the domicile of the purchaser. From the viewpoint of an American investor, it is convenient to express foreign currency as costing so many dollars.[6] Thus it is convenient to

[3]Goetzmann and Ukhov (2006).
[4]Lowenfeld (1907).
[5]Solnik (1976) and Levy and Sarnat (1970).

[6]Foreign currency exchange rates can be quoted in two ways. If an exchange rate is stated as the amount of dollars per unit of foreign currency, the exchange rate is quoted in direct (or American) terms. If the exchange rate is given as the amount of foreign currency per dollar, the quote is in indirect (or foreign) terms. The form of quotes differs across markets. In the interbank market, indirect quotes are used, whereas direct quotes are the norm in futures and options markets.

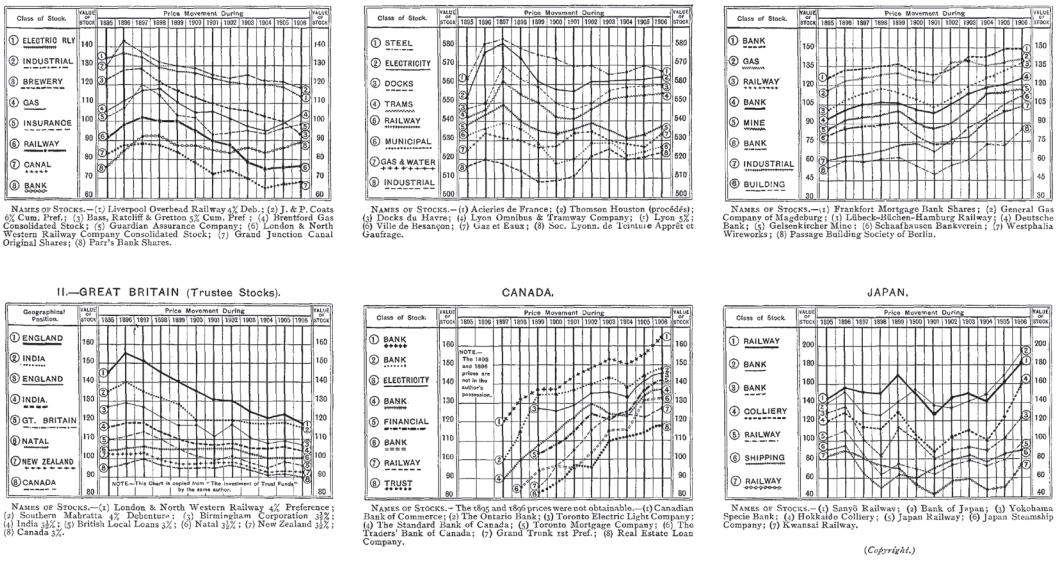
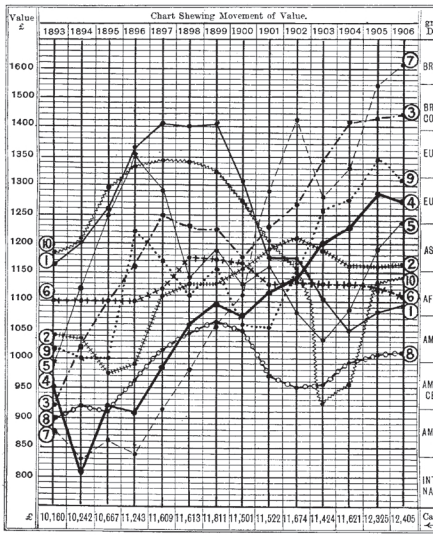


Figure 12.1a Typical price movement of the representative stocks of Great Britain, France, United States of America, Argentina. Produced in 1907.



Geographical Division	Quantities and Names of Stocks.	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902	1903	1904	1905	1906
		£	£	£	£	£	£	£	£	£	£	£	£	£	£
BRITISH.	① £700 London & North-Western Ord. Stock ..	37½	44	44½	50	50	50	50	44	38	42	41	40	42¾	44⅜
BRITISH COLONIES.	② £1 100 Canad. Pac. Ry. 4% Pf. Stk. (Non-Cm.)	44	44	44	44	44	44	44	44	44	44	44	44	44	44
EUROPE, NORTH.	③ 120 Ord. Shs. Deutsche Bank (Berlin)	48	54	60	60	60	63	66	66	66	66	60	72	72	72
EUROPE, SOUTH.	④ 90 Ord. "B" Shares Royal Sardinian Rly.	52	54	54	54	54	54	54	56	56	56	56	56	58½	58½
ASIA.	⑤ 180 Shares Bombay Gas Company ...	58½	58½	58½	58½	54	54	54	58½	58½	58½	58½	58½	63	63
AFRICA.	⑥ £1,100 Egyptian Delta Lt. Rly. 4% Deb. Stk.	44	44	44	44	44	44	44	44	44	44	44	44	44	44
AMERICA, NORTH.	⑦ 45 Shares Illinois Central Railroad Com....	45	45	45	45	45	45	54	54	54	54	54	63	63	
AMERICA, CENTRAL.	⑧ £700 Mexican Railway 6% Perp. Deb. Stock	42	42	42	42	42	42	42	42	42	42	42	42	42	
AMERICA, SOUTH.	⑨ £1,000 Buenos Ayres Wstrn. Rly. Ord. Stk.	60	52½	55	60	60	40	60	60	60	60	70	70	70	
INTERNATIONAL.	⑩ 130 Cm. Pf. Shs. Direct Spanish Telegraph...	65	65	65	65	65	65	65	65	65	65	65	65	65	
	Capital—Total Annual Values—Income £	495	503	512	523	518	501	524	534	528	532	535	546	564	566

Figure 12.1b Individual price movements of 10 stocks covering different geographical divisions. Produced in 1907.

express an exchange rate of 0.80 euro to the dollar, as the cost of 1 euro is $1.25. Assume the following information:

	1	2	
Time	Cost of 1 Euro	Value of German Shares	Value in Dollars (1 × 2)
0	$1.25	40 Euros	1.25 × 40 = $50
1	$1.00	45 Euros	1.00 × 45 = $45

Furthermore, assume that there are no dividends paid on the German shares. In this case the return to the German investor expressed in the home currency (euros) is

$$(1 + R_H) = \frac{45}{40} \qquad \text{or} \qquad R_H = 0.125 \text{ or } 12.5\%$$

However, the return to the U.S. investor is

$$(1 + R_{US}) = \frac{1.00 \times 45}{1.25 \times 40} = \frac{45}{50} \qquad \text{or} \qquad R_{US} = -0.10 \text{ or } -10\%$$

The German investor received a positive return, whereas the U.S. investor lost money because euros were worth less at time 1 than at time 0. It is convenient to divide the return to the American investor into a component due to return in the home or German market and the return due to exchange gains or losses. Letting R_x be the exchange return, we have

$$(1 + R_{US}) = (1 + R_x)(1 + R_H)$$

$$1 + R_x = \frac{1.00}{1.25} = 1 - 0.20 \qquad \text{or} \qquad R_x = -0.20$$

$$1 + R_H = \frac{45}{40} = 1 - 0.125 \qquad \text{or} \qquad R_H = 0.125$$

$$(1 + R_{US}) = (1 - 0.20)(1 + 0.125) = 1 - 0.10 \qquad \text{or} \qquad R_{US} = -0.10$$

Thus the 12.5% gain on the German investment was more than offset by the 20% loss on the change in the value of the euro. Restating the preceding equation,

$$(1 + R_{US}) = (1 + R_x)(1 + R_H)$$

Simplifying,

$$R_{US} = R_x + R_H + R_x R_H$$

In the example,

$$\begin{aligned} -0.10 &= -0.20 + 0.125 + (-0.20) \times (0.125) \\ &= -0.20 + 0.125 - 0.025 \end{aligned}$$

The last term (the cross-product term) will be much smaller than the other two terms, so that return to the U.S. investor is approximately the return of the security in its home market plus the exchange gain or loss. Using this approximation, we have the following expressions for expected return and standard deviation of return on a foreign security:

Expected return

$$\bar{R}_{US} = \bar{R}_x + \bar{R}_H$$

Standard deviation of return

$$\sigma_{US} = \left[\sigma_x^2 + \sigma_H^2 + 2\sigma_{Hx}\right]^{1/2}$$

As will be very clear when we examine real data, the standard deviation of the return on foreign securities (σ_{US}) is much less than the sum of the standard deviation of the return on the security in its home country (σ_H) plus the standard deviation of the exchange gains and losses (σ_x). This relationship results from two factors. First, there is very low correlation between exchange gains (or losses) and returns in a country (and therefore the last term, σ_{Hx}, is close to zero). Second, squaring the standard deviations, adding them, and then taking the square root of the sum is less than adding them directly. To see this, let

$$\sigma_x = 0.10$$
$$\sigma_H = 0.15$$
$$\rho_{Hx} = 0 \qquad \text{(to make the covariance zero)}$$

then

$$\sigma_{US}^2 = 0.10^2 + 0.15^2$$

and

$$\sigma_{US} = 0.18$$

Thus the standard deviation of the return expressed in dollars is considerably less than the sum of the standard deviation of the exchange gains and losses and the standard deviation of the return on the security in its home currency. The reader should be conscious of this difference in the tables that follow.

Having developed some preliminary relationships, it is useful to examine some actual data on risk and return.

THE RISK OF FOREIGN SECURITIES

Table 12.1 presents the correlation between the equity markets of several countries for the period 2002–2011. These correlation coefficients have been computed using monthly returns on market indexes. The indexes are computed by Morgan Stanley Capital International. They are market-weighted indexes with each stock's proportion in the index determined by its market value divided by the aggregate market value of all stocks in that market. All returns were converted to U.S. dollars at prevailing exchange rates before correlations were calculated. Thus Table 12.1 presents the correlation from the viewpoint of a U.S. investor. These are very low correlation coefficients relative to those found within a domestic market. The average correlation coefficient between a pair of U.S. common stocks is about 0.40, and the correlation between U.S. indexes is much higher. For example, the correlation between the S&P index of 500 large stocks and the rest of the stocks on the New York Stock Exchange is about 0.97. The correlation between a market-weighted portfolio of the 1,000 largest stocks in the U.S. market and a market-weighted portfolio of the next 2,000 largest stocks is approximately 0.92. Finally, the correlation coefficient between two 100-security portfolios drawn at random from the New York Stock Exchange is on the order of 0.95. The numbers in the table are much smaller than this, with the average correlation being 0.75.

The correlations shown in Table 12.1 are much higher than have appeared in any previous edition of this book, where average correlations were generally lower than 0.50. The reason for this is the crash of 2007–2008 was substantial, and felt worldwide. If this period

Table 12.1 Correlation Among Stock Indexes Measured in U.S. Dollars (2002–2011)

	Australia	Belgium	Brazil	Canada	China	Emerging Market	France	Germany	Hong Kong	Italy	Japan	Netherlands	Russia	Spain	Sweden	Switzerland	Taiwan	U.K.	U.S.
Australia	1.00																		
Belgium	0.78	1.00																	
Brazil	0.77	0.71	1.00																
Canada	0.84	0.74	0.78	1.00															
China	0.75	0.63	0.69	0.72	1.00														
Emerging market	0.89	0.80	0.83	0.87	0.82	1.00													
France	0.84	0.88	0.73	0.79	0.67	0.84	1.00												
Germany	0.80	0.83	0.71	0.75	0.66	0.83	0.95	1.00											
Hong Kong	0.73	0.67	0.68	0.76	0.83	0.83	0.69	0.66	1.00										
Italy	0.80	0.85	0.65	0.76	0.61	0.79	0.94	0.90	0.67	1.00									
Japan	0.61	0.56	0.50	0.62	0.52	0.65	0.57	0.56	0.55	0.56	1.00								
Netherlands	0.82	0.89	0.75	0.77	0.66	0.84	0.94	0.92	0.71	0.89	0.59	1.00							
Russia	0.72	0.63	0.65	0.76	0.56	0.79	0.66	0.62	0.65	0.65	0.56	0.63	1.00						
Spain	0.78	0.81	0.69	0.69	0.64	0.78	0.91	0.86	0.64	0.90	0.54	0.86	0.56	1.00					
Sweden	0.80	0.82	0.71	0.73	0.67	0.83	0.89	0.88	0.72	0.83	0.55	0.89	0.59	0.81	1.00				
Switzerland	0.79	0.81	0.65	0.69	0.61	0.77	0.87	0.86	0.65	0.82	0.61	0.85	0.60	0.78	0.81	1.00			
Taiwan	0.72	0.66	0.61	0.69	0.69	0.83	0.68	0.68	0.75	0.62	0.50	0.71	0.61	0.62	0.69	0.62	1.00		
U.K.	0.86	0.87	0.76	0.83	0.70	0.85	0.91	0.87	0.75	0.89	0.61	0.90	0.71	0.84	0.86	0.84	0.68	1.00	
U.S.	0.86	0.82	0.74	0.85	0.65	0.86	0.90	0.89	0.71	0.84	0.61	0.89	0.66	0.80	0.86	0.82	0.70	0.90	1.00
			Average correlation coefficient				0.75												

was excluded from the data, the correlation coefficient would be closer to the historical levels. However, this gives rise to several important issues. The major justification for international diversification is that historically this has led to lower risk portfolios. The lower risk has been primarily due to the low correlation between domestic and foreign portfolios. Examining earlier editions of the book shows that average correlations have risen. A principal reason for this is the European monetary union has eliminated exchange rate fluctuations between member countries and markets have become more integrated. A second reason is what happens in crashes. Recent evidence shows that in financial crises, correlation in equity returns goes up and that the risk reduction properties of international diversification are reduced and possibly eliminated in these crashes.

Risk depends not only on correlation coefficients but also on the standard deviation of return. Table 12.2 shows the standard deviation of return for an investment in the common equity indexes. It should be emphasized once again that the standard deviation is calculated on market indexes and is therefore a measure of risk for a well-diversified portfolio, consisting only of securities traded within the country under examination.

As shown in the last section, there are two sources of risks. The return on an investment in foreign securities varies because of variation of security prices within the securities home market and because of exchange gains and losses. Note that in some cases the total risk is less than the domestic risk. The reduction in correlation when exchange rates are taken into account comes about because for these countries in this period exchange fluctuations were negatively correlated with movements in the local market.

The column headed "Domestic Risk" in Table 12.2 shows the standard deviation of return when returns are calculated in the indexes' own currency. Thus the standard deviation of 18.13% for Japan is the standard deviation when returns on Japanese stocks are calculated in yen. The second source of risk is exchange risk. Exchange risk arises because the exchange rate between the yen and dollar changes over time, affecting the return to a U.S. investor on an investment in Japanese securities. The variability of the exchange rate for each currency

Table 12.2 Risk for U.S. Investors in Stocks 2002–2011

Country	Domestic Risk	Exchange Risk	Total Risk
Australia	13.41	13.87	23.38
Belgium	21.80	11.03	25.86
Brazil	24.20	19.94	36.80
Canada	14.64	10.04	21.94
China	31.75	1.50	28.21
Emerging market	19.16		24.95
France	18.78	11.03	23.44
Germany	23.13	11.03	27.07
Hong Kong	21.90	0.53	21.94
Italy	19.88	11.03	24.90
Japan	18.13	9.13	17.16
Netherlands	20.67	11.03	24.09
Russia	32.74	9.52	35.70
Spain	21.12	11.03	26.11
Sweden	22.30	12.70	27.74
Switzerland	14.89	11.94	17.46
Taiwan	22.88	5.20	25.77
U.K.	15.74	9.32	18.32
U.S.	16.19	0.00	16.19

converted to dollars is shown in the column titled "Exchange Risk." As discussed in the last section, the exchange risk and the within-country risk are usually relatively independent (in this period they were negatively correlated) for many countries, and standard deviations are not additive. Thus total risk to the U.S. investor is much less than the sum of exchange risk and within-country risk. For example, the standard deviation of Japanese stocks in yen is 18.13%. The standard deviation of changes in the yen dollar exchange rate is 9.13%. However, the risk of Japanese stocks in dollars when both fluctuations are taken into account is 17.16%. It should be emphasized that the variability of exchange rates is calculated by examining the variability of each currency in dollars. Thus the total risk is measured from a U.S. investor's point of view.

As shown in Table 12.2 over the 2002–2011 time period, the standard deviation of an index of the U.S. equity market was low relative to the standard deviation of most other market indexes when the standard deviation of returns was calculated in its own currency (domestic risk). When the effect of exchange risk is taken into account, the higher risk of foreign markets was even more pronounced. These results are not atypical. Solnik (1988), Kaplanis and Schaefer (1991), and Eun and Resnick (1988) find the same results for different periods. We found the same results in all earlier editions of this book.

The risk of the portfolio depends on the correlation between markets and the standard deviation of each market. In this period there was little risk reduction through international diversification. This is due to the crash in 2007–2008, which caused higher-than-normal exchange rate fluctuations, and principally the higher than normal correlation coefficients. To understand if this is atypical, we need to better understand what determines the magnitude of the correlation.

Equity market correlations have changed dramatically over the long term. This variation in correlation affects the benefits of international investing and matters a lot to investors. Goetzmann, Li, and Rouwenhorst (2005) looked at the correlation of world stock markets during different periods in world history from 1875 to 2000.[7] Figure 12.2 shows that correlations reached a high point in the Great Depression, but this peak was nearly matched by the end of the sample in the year 2000. In fact, correlations continued to increase even more in the first decade of the twenty-first century.

Dividing history into subperiods based on levels of market integration, that is low versus high barriers to trade, the authors found that correlations were highest when barriers to cross-border flows were lowest. In other words, during periods of globalization, international diversification delivered less reduction in risk. Quinn and Voth (2008) take this analysis a step further and attribute the variation in equity market correlations to international capital account openness, and in related work, Bekaert and Harvey (2000) show that when stock markets open up to foreign investment, their correlation with the world market portfolio increases.[8] As a result, the largest benefit to international diversification is likely to be in the markets most difficult to access. Liberalization is a two-edged sword. It provides access but reduces benefit.

An analytical measure of the benefits of international diversification can be examined by assuming equal investment in each country. As shown in Chapter 4, the risk of an equally weighted portfolio in n countries and using upper bars to indicate averages is

$$\frac{1}{n}+\left(\frac{n-1}{n}\right)\times\frac{\overline{\text{Cov}\left(x_i,x_j\right)}}{\overline{\text{Var}\left(x_i\right)}}$$

[7]Goetzmann, Li, and Rouwenhorst (2005).
[8]Quinn and Voth (2008) and Bekaert and Harvey (2000).

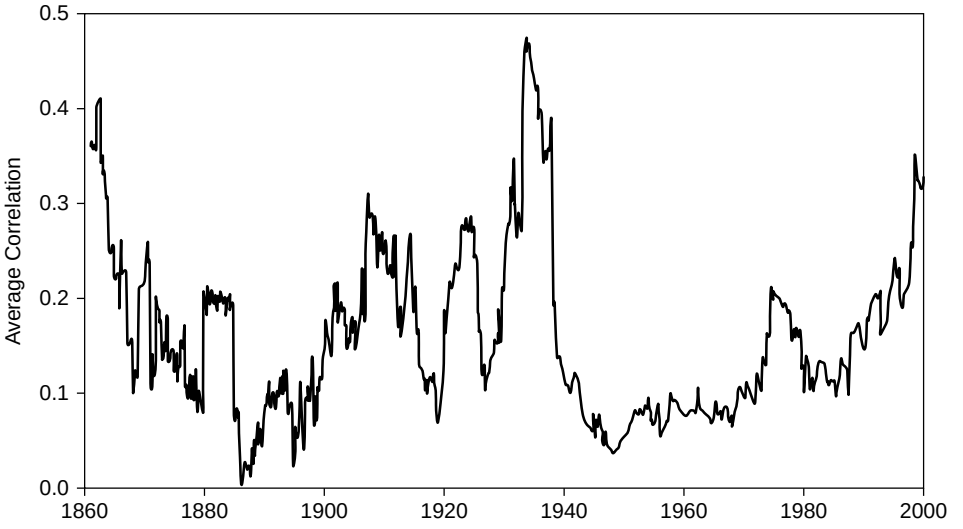


Figure 12.2 Average correlation of capital appreciation returns for all available markets. This figure shows the time series of the average off-diagonal correlation of dollar-valued capital appreciation returns for all available markets. A rolling window of 60 months is used. *Source*: Goetzmann, Li, and Rouwenhorst (2005)

Using the formula, we can see the risk reduction in various periods. Figure 12.3 from Goetzmann, Li, and Rouwenhorst (2005) shows that the risk reduction of holding a portfolio of 21 international stock markets in the period 1940–1945 was twice as great as in the period 1972–2000.

When we examine the risk reduction of international investment, we can see that risk reduction is due to two factors: the average covariance of the markets and the number of

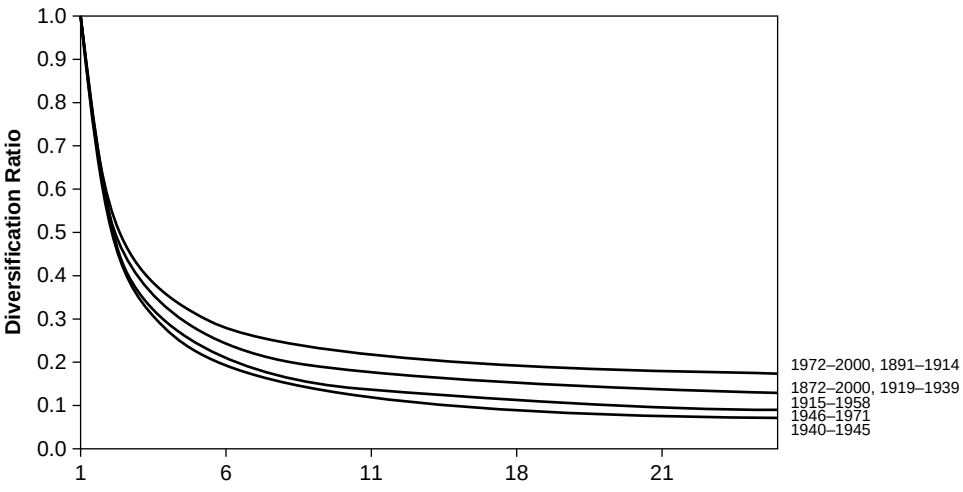


Figure 12.3 Risk reduction from international diversification: Selected periods. This figure shows the ratio of the average covariance of the equally weighted portfolio of country indexes scaled by the average variance of the country indexes, as a function of the number of countries in the portfolio. *Source*: Goetzmann, Li, and Rouwenhorst (2005)

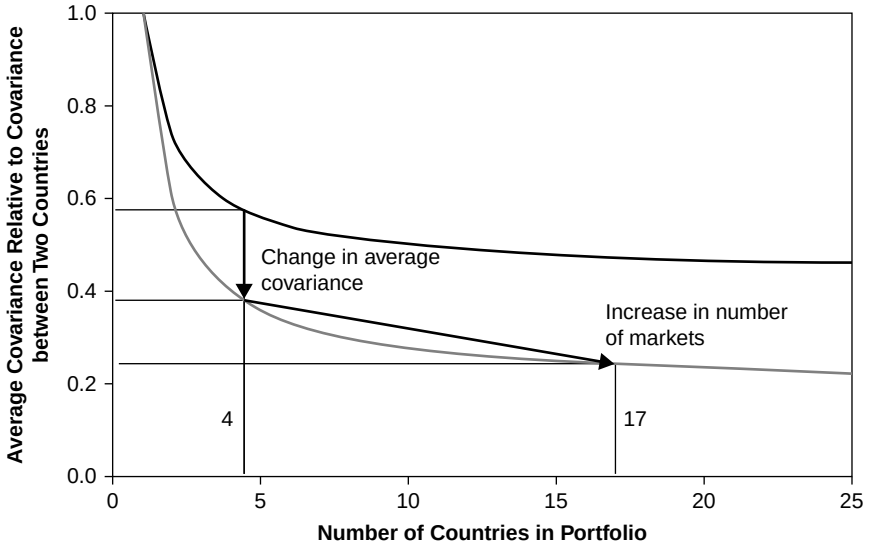


Figure 12.4 Average relative covariance versus investment opportunity set. *Source*: Elton, Gruber, Brown, and Goetzmann (2014).

markets available. In periods of globalization, *n*, the number of markets grows as barriers to international investing fall, even while the average covariance of markets increases. (See Figure 12.4.)

Figure 12.5 compares the risk reduction of two international investment strategies during the period 1975 to 2000. The first invests equally in equity indexes of four major countries:

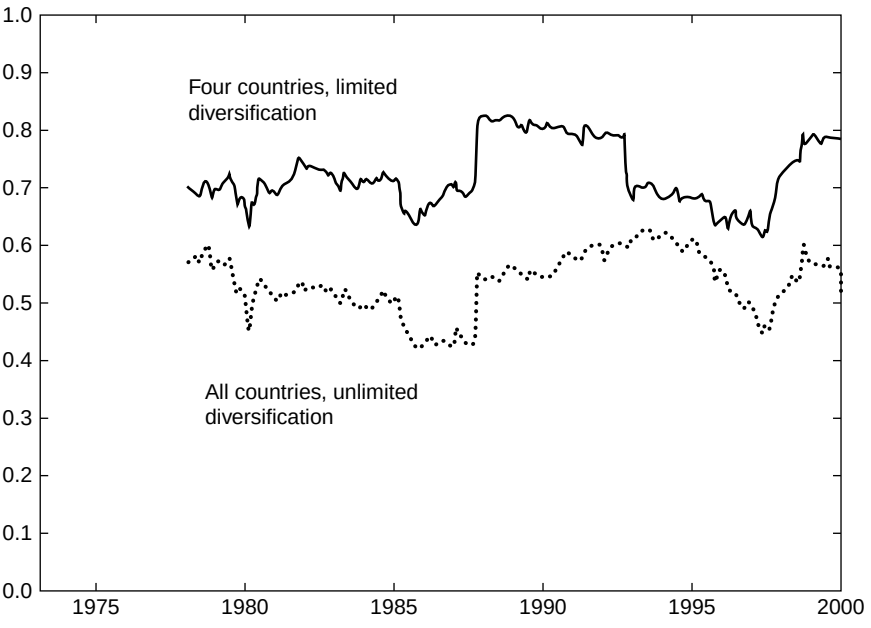


Figure 12.5 Diversification with capital market weights. This figure shows the risk reduction for portfolios of 45 country indexes and the risk reduction of the four core countries. A rolling window of 120 months is used. Returns are exponentially weighted with a half time of 60 months. *Source*: Elton, Gruber, Brown, and Goetzmann (2014).

the United States, the United Kingdom, France, and Germany. The second invests equally in 45 country portfolios. Notice the substantial benefits of extended diversification given by the second strategy. Even though correlations were high during the period, investing in a broadly diversified global equity portfolio still delivered a reduction in risk compared to the home country of 50%.

Although barriers to international capital flows are one determinant of the covariance structure of global markets, Richard Roll (1992) suggested that industrial differences are also potentially important and showed evidence that industrial specialization matters. Countries that produce natural resources, for instance, might have a low correlation to countries that are industrial producers. Heston and Rouwenhorst (1994) tested this proposition using an extensive dataset from 12 European countries from 1978 to 1992, prior to the adoption of the euro. After decomposing index returns into industry and country effects, they found that the benefits of diversification across countries, even within a single industry, outweighed the benefits of diversification across all industries within a single country.[9] Their findings using data from the modern period of globalization echoed those of Lowenfeld, working with data from nearly a century earlier.

Heston and Rouwenhorst concluded that institutional differences across countries must play a major role in explaining correlation and volatility. This finding is consistent with the documented time-varying benefits of international investing, depending on the degree of market openness to international flows. This notion of market "openness" is an important one in international investing.

MARKET INTEGRATION

A more formal expression of the notion of financial market openness is the concept of market integration. A world with strict barriers to cross-border investing is called *segmented* as opposed to *integrated*. For example, the Chinese stock market maintains separate classes of shares: one class for domestic investors and one for foreign investors. These two types of shares are not perfectly correlated because the forces of arbitrage cannot operate: investors cannot "comparison shop" across the two different markets and force prices to align. Because the two share classes represent claims on the same economic benefits, if the two markets were perfectly integrated, the price of the foreign and domestic Chinese shares should be equal. This is also called the *law of one price*.[10] Markets are said to be integrated if economically identical claims in two markets have the same price.

An interesting test of the hypothesis of global stock market integration is the case of Royal Dutch and Shell.[11] For many years shares of Shell traded mainly in the United Kingdom, and shares of Royal Dutch traded in the Netherlands and the United States. Despite this, the two types of shares were equivalent claims on the same oil firm, and if the U.K., Dutch, and U.S. markets were perfectly integrated, their prices should have remained in strict proportion based on their relative claims to the firm. Two empirical studies of the Royal Dutch and Shell share prices showed that they violated the law of one price—deviating from parity as much as 40% at times.[12] This deviation could not have been caused by industrial differences, and thus could only be due to local market factors. An important point made by both studies is that, during the period when these stocks traded, it was not easy to take advantage of the large price discrepancies. An investor buying the "cheap" stock and selling the "expensive" stock would have to hold that position

[9]Heston and Rouwenhorst (1994).

[10]See Chen and Knez (1995).

[11]Reprinted from *The Journal of Financial Economics* **53**, No. 2, "How are stock prices affected by the location of trade?" 189–216, 1999, with permission from Elsevier.

[12]Froot and Dabora (1999) and Rosenthal and Young (1990).

until all future dividends were paid to profit. Thus three of the major markets in the world at one time may not have been perfectly integrated, but exploiting the market segmentation was also not straightforward. Froot and Dabora study the deviations of Royal Dutch and Shell from the law of one price over time. They find that the largest deviations (25% to over 40%) occurred in the 1980s, and by the mid-1990s deviations from the law of one price were much smaller (about 5%). This suggests that the capital markets may have become more integrated over the period of the author's study.

The major justification for international diversification is the low correlation among markets. As we have seen, market integration is a major factor affecting the size of the correlation. The greater integration of the European markets may mean that investors will have to invest in non-European markets to get the benefits of international diversification. Before leaving this section, we need to examine returns in various markets.

RETURNS FROM INTERNATIONAL DIVERSIFICATION

The period from 2002 through 2011 was not an especially favorable time for U.S. markets relative to foreign markets. Table 12.3 shows the average annual returns from January 2002 to December 2011 on several international markets. The "Exchange Gain" column is the difference between the return in the assets home country and the assets return in the United States.[13]

The column in Table 12.3 that presents returns in U.S. dollars shows every country had returns above the United States. Thus most internationally diversified equity portfolios would have had a higher return than the U.S. market index over this period. During this period, international diversification had the advantage of larger average returns.

Table 12.3 Return to U.S. Investor in Stocks 2002–2011 (percent per annum)

Name	Own Country	Exchange Gain	To U.S. Investor
Australia	6.91	8.87	15.79
Belgium	6.91	−1.23	5.68
Brazil	7.29	18.43	25.72
Canada	6.41	7.20	13.62
China	6.32	11.88	18.21
Emerging market	5.91	13.00	18.91
France	5.25	0.99	6.24
Germany	5.60	3.16	8.75
Hong Kong	5.76	4.96	10.72
Italy	5.21	−0.91	4.30
Japan	5.94	−1.40	4.54
Netherlands	6.87	−0.17	6.70
Russia	7.66	12.14	19.81
Spain	5.46	5.45	10.90
Sweden	5.34	7.76	13.11
Switzerland	4.64	4.37	9.00
Taiwan	5.54	2.48	8.02
U.K.	5.20	1.19	6.39
U.S.	5.55	0	5.55

[13]Earlier we showed that the expected return to a U.S. investor is not the sum of exchange gains and losses and the return in the investor's home country. Thus column 2 includes not only the exchange return but also includes all joint effects of the country and exchange return.

Although these results are appropriate for the period discussed, it is useful to examine other periods. Solnik (1988) studied equity indexes for 17 countries for the years 1971–1985. For all but two countries the return on the foreign index expressed in dollars was greater than the return on the U.S. equity index. The exchange gain from holding foreign equities added 0.2% on average to this return. No country had a lower return when return was expressed in U.S. dollars.

For portfolio decisions, estimates of future values of mean return, standard deviation, and correlation coefficients are needed. The correlation coefficients between international markets have been very low historically relative to intracountry correlations. As Europe has integrated its markets and as all countries have moved toward greater integration, these coefficients have risen.[14] However, they are still likely to be low relative to intracountry correlation. For example, the correlation coefficient among countries whose economies are relatively highly integrated, such as Canada and the United States, or the Scandinavian countries, is still much lower than the intracountry correlation coefficients. Thus international diversification is likely to continue to lead to risk reduction in the foreseeable future. However, we know of no economic reason to argue that returns in foreign markets will be higher or lower than for domestic markets.

THE EFFECT OF EXCHANGE RISK

Earlier we showed how the return on a foreign investment could be split into the return in the security's home market and the return from changes in exchange rates. In each of the prior tables, we separated out the effect of changes in the exchange rate on return and risk. In Tables 12.2 and 12.3, the columns entitled "Exchange Risk" or "Exchange Gain" calculated the effect of converting all currencies into dollars. Obviously if we were presenting the same tables from a French point of view, the exchange "Expected Return" and "Risk" columns would be different, because they would contain results as if all currencies were converted to euros. Because euros have not fluctuated perfectly with the dollar, these columns would be different. Thus the country of domicile affects the expected returns and risk (including correlation coefficients) from international diversification.

Table 12.4 illustrates this by computing expected return and risk from the U.S. investor's point of view (which is a repeat of prior tables) and from the euro point of view for a country whose currency is euros. The numbers are clearly quite different. It is possible to protect partially against exchange rate fluctuations. An investor can enter into a contract for future delivery of a currency at a price that is fixed now. For example, an American investor purchasing

Table 12.4 The Effect of Country of Domicile on Mean Returns and Risk 2002–2012

	In Euros		In Dollars	
Country	Mean Return	Standard Deviation	Mean Return	Standard Deviation
China	9.91	30.02	18.21	28.21
France	1.45	18.78	6.24	28.21
Russia	14.54	32.46	19.81	35.7
Switzerland	3.93	13.94	9	17.46
U.K.	2.09	15.38	6.39	18.32
U.S.	−0.19	15.46	5.55	16.19

[14]In particular, exchange rates between European union currencies are fixed. Although European union currencies will continue to fluctuate with the U.S. currency, any advantage in diversifying across these currencies will be eliminated.

German securities could simultaneously agree to convert euros into dollars at a future date and at a known rate. If the investor knew exactly what the security would be worth at the end of the period, he or she would be completely protected against exchange rate fluctuations by agreeing to switch an amount of euros exactly equal to the value of the investment. However, given that, in general, the end-of-period value of the investment is random, the best the investor can do is protect against a particular outcome (e.g., its expected value).[15]

As shown earlier, the standard deviation of foreign investments generally increases as a result of exchange risk. If exchange risk was completely hedged, then the "Domestic Risk" column in Tables 12.2 and 12.3 would be the relevant column used to measure risk.

Although we will not present the tables, the correlation coefficients are generally lower when we calculate the correlation between returns assuming exchange risk is fully hedged away. Exchange movement generally increase the correlation among countries' returns.

The effect on expected return is less clear. Figure 12.3 shows that during the 2002–2011 period, exchange movements caused gains to U.S. investors for most countries. The same table in the 1990s would have shown mostly losses. Also, the gain to the U.S. investor is the loss to the foreign investor, so that a different table would hold if we expressed returns in, for example, euros. Thus the effect of eliminating exchange gains or losses on expected return varies from country to country and period to period.

One way to determine whether international diversification will be a useful strategy in the future is to analyze how low expected returns in foreign countries would have to be for an investor not to gain via international diversification.

RETURN EXPECTATIONS AND PORTFOLIO PERFORMANCE

Most of the literature on domestic and international diversification tells us that history is a much better guide in forecasting risk than it is in forecasting returns. If we accept the historical data on risk as indicative of the future, for any assumed return on the U.S. market we can solve for the minimum return that must be offered by any foreign market to make it an attractive investment.

Hold foreign securities as long as[16]

$$\frac{\bar{R}_N - R_F}{\sigma_N} > \frac{\bar{R}_D - R_F}{\sigma_D}\rho_{N.D} \tag{12.1}$$

where

$\bar{R}_N$ is the expected return on the foreign securities in dollars

$\bar{R}_D$ is the expected return on domestic securities

σ_N is the standard deviation of the foreign securities in dollars

[15]Procedures exist for changing the hedge through time to eliminate most of the exchange risk. See Kaplanis and Schaefer (1991).

[16]From Chapter 4 the first-order conditions are

$$\bar{R}_N - R_F = Z_N\sigma_N^2 + Z_D\rho_{N.D}\sigma_D\sigma_N$$
$$\bar{R}_D - R_F = Z_N\rho_{N.D}\sigma_D\sigma_N + Z_D\sigma_D^2$$

Setting Z_N equal to zero and eliminating Z_D results in the preceding equation as an equality. Increasing $\bar{R}_N$ would cause Z_N to be greater than zero. For a more detailed derivation see Elton, Gruber, and Rentzler (1987).

This analysis assumes foreign securities cannot be shorted. If they can be shorted, then markets for which Equation (12.1) does not hold are candidates for short sales.

σ_D is the standard deviation of domestic securities

$\rho_{N.D}$ is the correlation between domestic and foreign securities

R_F is the risk-free rate of interest

If we rearrange the expression (12.1), we hold foreign securities as long as[17]

$$\overline{R}_N - R_F > \left[\overline{R}_D - R_F\right]\left[\frac{\sigma_N \rho_{N.D}}{\sigma_D}\right] \tag{12.2}$$

As long as the expression in the last bracket is less than 1, foreign securities should be held even with expected returns lower than those found in the domestic market.

What is foreign to one investor is domestic to another, however. Are there any circumstances where international diversification does not pay for investors of all countries?

To understand this issue, consider the U.S. and Japanese markets. If holding the two markets lowers risk, and given the numbers in the prior tables, it does, then if investors in the two markets agree on expected returns, we have one of three situations: both gain from diversification, the U.S. investor gains, or the Japanese investor gains. In all three cases, however, at least one investor should diversify internationally. If the investors do not agree on returns in the two markets, then it is possible that neither the U.S. investor nor the Japanese investor will benefit from international diversification. For example, assume U.S. investors believe that Japanese markets have an expected return of 5%, whereas U.S. markets would have an expected return of 10%. Further assume that Japanese investors believe Japanese markets have an expected return of 10%, whereas U.S. markets have an expected return of 5%. Under this set of expected returns, neither U.S. nor Japanese investors would wish to diversify internationally. Are there any circumstances where investors in all countries could rationally believe that returns are higher in their country relative to the rest of the world? The answer is *yes*!

If governments tax foreign investments at rates very different from domestic investments, then the pattern just discussed would be possible for after-tax returns. Differential taxation has occurred in the past, continues to occur today, and will likely persist into the future. Second, many countries impose a withholding tax on dividends. Taxable investors may receive a domestic credit for the foreign tax withheld and thus not have lowered returns. However, for nontaxable investors (or for a nontaxable part of an investor's portfolio such as pension assets), the withholding is a cost that lowers the return of foreign investment. A third situation that could cause foreign investments to have a lower return than domestic investments for all investors is if there were differential transaction costs for domestic and foreign purchases. This could occur if there was difficulty in purchasing foreign securities or currency controls existed. For example, there may be restrictions in converting domestic to foreign currency that could affect returns. The exchange of currency A for B might take place at an official rate higher than the free market rate, and there might be an expectation of a later reversal. A fourth situation that can result in investors in all countries having an expectation of higher returns from domestic investments relative to foreign is a danger of a government restricting the ability of foreigners to withdraw funds. Governments can and do place such restrictions on foreigners, and this can reduce returns to foreigners. The considerations just discussed are real and can affect the returns from international diversification.

[17]Multiplying the numerator and denominator of the expression in the brackets by σ_D shows that the expression in the brackets is the beta of the non-U.S. markets on the U.S. index.

Before leaving this section, one other issue needs to be discussed. It has been suggested that investors could confine themselves to a national market and receive most of the benefits of international diversification by purchasing stocks in multinational corporations. Jacquillat and Solnik (1978) have tested this for the American investor. They found that stock prices of multinational firms do not seem to be affected by foreign factors and behave much like the stocks of domestic firms. The American investor cannot gain much of the advantage of international diversification by investing in the securities of the multinational firm.

EMERGING MARKETS

One of the major effects of globalization of capital markets in the last few decades has been the emergence of new capital markets in many countries. The introduction of equity markets in China and Russia in 1990, the opening of Eastern Europe, the founding of markets in Africa and Asia, as well as the general revival and growth of equity markets through the latter part of the twentieth century opened up considerable new opportunities for international investing. The term "emerging markets" was coined by the World Bank to refer to these new exchanges, and over the past 20 years, the International Financial Corporation (IFC) has maintained indexes for many of them. Early studies of emerging markets using IFC data showed them to be high return but also high risk, although the evidence on high return depended to some extent on the time period over which data are measured. Figure 12.6 shows the performance of the S&P/ICF emerging market index over the period from 1989 to 2011. Emerging markets outperform the U.S. index over the entire period, although most of the outperformance came after the year 2000. Barry, Peavy, and Rodriguez (1998) extend the emerging market index back to 1975 and find lower relative performance prior to 1989 as well. Some of the largest economies in the world, such as Brazil, Russia, India, and China (terms BRICS), boomed in the post-1989 period and opened up to foreign investment. The equity markets of these countries allowed investors to participate in their growth. However, the high performance of the BRICs may be due to an unusual episode in global capital markets rather than being indicative of future higher returns.

A more subtle factor to consider in emerging market investing is selection bias. The IFC historically collected data for markets that are recently successful. Jorion and Goetzmann (1999) noted that most emerging markets actually have a long history, interrupted by wars and other adverse events, and that emerging markets are typically "re-emerging." It is difficult to get information about stocks in countries that did not recover from bad times. The recent growth and integration of re-emerging markets into the world capital markets may therefore be temporary—a result of world market liberalization that is reversed in periods of global distress.[18] Figure 12.7 is a useful example. It takes data on seven stock market indexes collected originally by the League of Nations and then by the United Nations on world stock markets. Notice that four of the markets grew significantly over much of the twentieth century, surviving the Great Depression, World War II, and the global economic malaise of the 1970s; however, three countries faired poorly. The data for Argentina stopped in the late 1950s, the data for Chile has a break in the 1970s corresponding with a political change, and the Columbian stock market, while continuous, has lost value steadily since a peak in the 1940s. All three countries since 1990 have been regarded as emerging markets. The challenge for investors is to decide whether their recent performance or their long-term historical performance is a better indicator of their future.

[18]Goetzmann and Jorion (1999) and Bekaert and Harvey (1995).

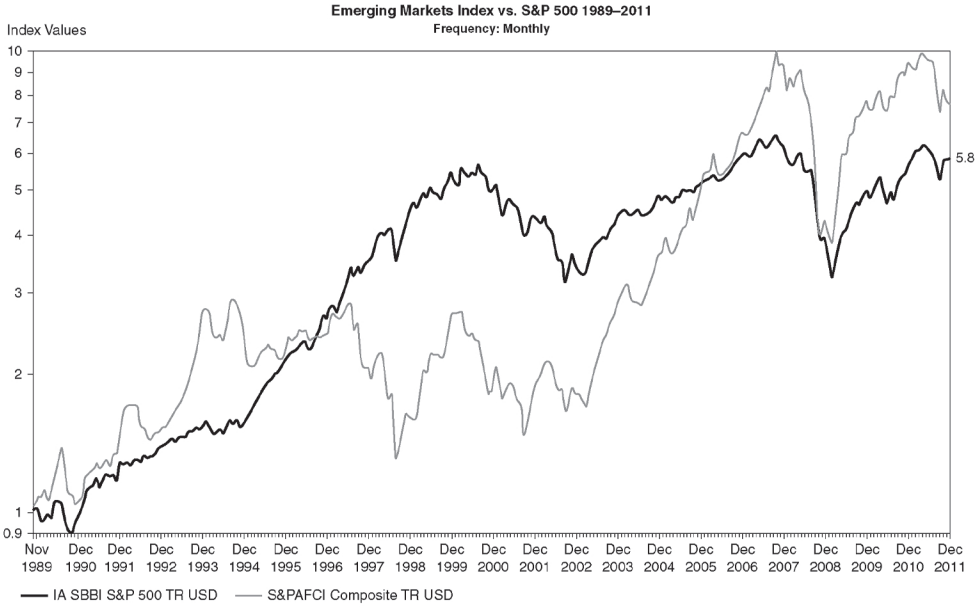


Figure 12.6 S&P/IFC Emerging Market Composite Index vs. S&P 500 Index, 1989–2011. (Courtesy Morningstar/Ibbotson Encorr)

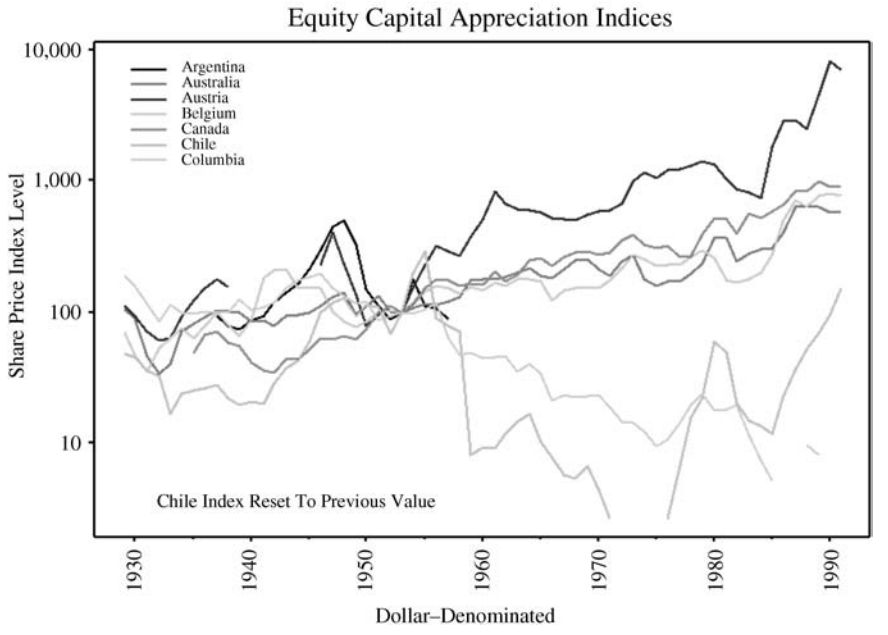


Figure 12.7 Sample of global markets from 1922 to 1994. *Source*: Goetzmann and Jorion (1991).

Baekert and Harvey (1995) explicitly study how the expected return characteristics of emerging markets change through time. They document the time-varying nature of world capital market integration and the potential for emerging market returns to decline after emergence. Because emerging markets are by definition ones that have recently grown—sometimes dramatically—ex post studies of emerging markets will likely show high returns and time variation in performance. It may not be wise to extrapolate the prior returns of high-growth markets that emerged in times of global market integration.

International Diversification of Bonds

Table 12.5 shows the correlation between the bond indexes of 13 countries for the years 2002–2011. These indexes are value-weighted indexes of the major issues in each country. Once again the correlations are very low relative to the correlations of two intracountry indexes or bond portfolios. The average correlation between countries shown in Table 12.5 is 0.51. In contrast, Kaplanis and Schaefer (1991) show an average correlation between countries of 0.43 for long-term bond indexes in their sample period, and Chollerton, Pieraerts, and Solnik (1986) find 0.43. This can be contrasted with the correlation between two typical American bond mutual funds of 0.94 and the correlation between the U.S. government and corporate bond index of 0.98.

As shown in Table 12.6, for long-term bonds, the standard deviation of the U.S. bond index is low compared to the standard deviation of each index calculated in its own currency. When returns are adjusted for changes in exchange rates and all returns are expressed in dollars, the risk for the U.S. bond index is much lower than for any foreign index. This illustrates the importance of exchange rate fluctuations on returns and risk.

Table 12.5 Correlations Among Bond Indicies Measured in U.S. Dollars (2002–2011)

Country	Australia	Belgium	Brazil	Canada	Euro	France	Japan	Netherlands	Russia	Spain	Sweden	U.K.	U.S.
Australia	1.00												
Belgium	0.74	1.00											
Brazil	0.40	0.21	1.00										
Canada	0.75	0.59	0.39	1.00									
Euro	0.76	0.99	0.22	0.59	1.00								
France	0.76	0.98	0.22	0.59	0.99	1.00							
Japan	0.19	0.39	0.09	0.11	0.39	0.42	1.00						
Netherlands	0.78	0.98	0.22	0.60	0.99	1.00	0.41	1.00					
Russia	0.65	0.45	0.51	0.62	0.46	0.44	0.01	0.46	1.00				
Spain	0.62	0.85	0.17	0.49	0.83	0.81	0.29	0.81	0.31	1.00			
Sweden	0.76	0.86	0.23	0.60	0.88	0.88	0.35	0.89	0.48	0.69	1.00		
U.K.	0.61	0.68	0.04	0.47	0.69	0.71	0.25	0.73	0.37	0.48	0.67	1.00	
U.S.	0.42	0.53	0.28	0.43	0.53	0.56	0.42	0.56	0.52	0.39	0.43	0.47	1.00
	Average correlation coefficient				0.51								

Table 12.6 Risk for U.S. Investors in Bonds 2002–2012

	Domestic Risk	Exchange Risk	Total Risk
Australia	3.82	13.87	13.08
Belgium	3.84	11.03	11.91
Canada	3.97	10.04	10.03
EUR	4.17	11.03	11.91
France	3.84	11.03	11.55
Japan	2.17	9.13	9.83
Netherlands	3.72	11.03	11.41
Spain	4.66	11.03	16.60
Sweden	4	12.7	12.42
U.K.	5.32	9.32	10.00
U.S.	3.73	0.00	3.73

Table 12.7 shows the return on various bond indexes over the period 2002–2012. In this period, the return to U.S. investors was higher than returns in most other countries when all returns are calculated in the home currency. However, because the dollar depreciated substantially when exchange gains and losses were taken into account, foreign investment had much higher returns.

OTHER EVIDENCE ON INTERNATIONALLY DIVERSIFIED PORTFOLIOS

In prior sections we have presented the considerations that are important in deciding on the reasonableness of international diversification. Obviously, we feel that the type of analysis we have presented is the relevant way to analyze the problem. However, several studies analyze the reasonableness of international diversification by examining the characteristics of international portfolios selected using historical data. The most common approach attempts to show the advantages of international diversification by forming an optimal portfolio of international and domestic securities using historical data and comparing the return to an exclusively domestically held portfolio over the same time period. It should not surprise the reader that knowing the exact values of mean returns, variance,

Table 12.7 Return to U.S. Investors in Bonds 2002–2012 (percent per annum)

	Domestic Return	Exchange Return	Total Return
Australia	6.47	7.86	14.33
Belgium	4.89	4.74	9.63
Canada	6.09	5.11	11.20
EUR	4.03	5.31	9.34
France	5.14	4.75	9.89
Japan	1.66	6.19	7.85
Netherlands	5.62	4.73	10.35
Spain	4.42	7.84	12.26
Sweden	5.87	5.06	10.93
U.K.	6.6	1.30	7.90
U.S.	5.70	0.00	5.70

and covariances for international markets allows construction of portfolios that dominate investment exclusively in the domestic portfolio. A variant of this analysis presents the efficient frontier using historical data with and without international securities and "shows" that adding international securities improves the efficient frontier.

Although examining historical data is interesting, the real of test of international diversification is the performance of funds that hold internationally diversified portfolios. Table 12.8 shows data for the 23 international funds (funds that invest only in international securities) that existed over the last decade together with data on the S&P index.

The major promise of international diversification is the low correlation between domestic securities and foreign securities. As shown in Table 12.7, the average correlation between the fund return and the S&P index was 0.92. These correlations are somewhat higher than the correlations between the international stock indexes and the U.S. indexes presented in Table 12.1.

The column titled "beta" shows the responsiveness of international funds to a change in the S&P index. The beta is the beta introduced in Chapter 5, where we discussed the single-index model. For the 23 funds the average beta is 1.09. Both of these numbers are unusually high and reflect the crash of 2007–2008. In early editions of this book the numbers were much lower and were furthermore much lower than what we observe on domestic funds. The average standard deviation of an international portfolio was also somewhat higher than the S&P index.

Table 12.8 Performance Data on Stock Funds (2002–2012)

Fund	Mean Monthly Return	Standard Deviation	Beta	Correlation with CRSP Index
AllianceBern International A	0.35	5.75	1.13	0.92
Artio International Equity A	0.58	5.43	1.03	0.89
Columbia Acorn International Z	0.98	5.70	1.07	0.88
Eagle International Equity A	0.40	5.56	1.08	0.90
BlackRock International lnv A	0.44	5.66	1.11	0.92
Consulting Group International Eq Invst	0.54	5.60	1.12	0.93
Frost International Equity A	0.75	5.18	1.01	0.92
Fidelity Advisor Diversified Intl A	0.53	5.48	1.07	0.91
Glenmede International	0.55	6.01	1.18	0.92
Huntington International Equity A	0.59	5.10	1.00	0.92
Ivy International Core Equity A	0.70	5.56	1.08	0.91
Legg Mason Batterymarch Intl Eq A	0.48	5.40	1.05	0.91
MFS Research International A	0.61	5.38	1.06	0.92
UBS International Equity A	0.43	5.62	1.13	0.94
New Century International	0.69	5.60	1.11	0.92
Prudential International Equity A	0.43	5.86	1.17	0.93
T. Rowe Price International Stock Fd	0.53	5.96	1.19	0.93
PACE International Equity A	0.44	5.53	1.10	0.93
Saratoga International Equity A	0.23	5.73	1.14	0.93
State Farm International Equity A	0.40	5.76	1.15	0.93
Thomas White International A	0.77	5.66	1.10	0.91
Vantagepoint International	0.48	5.34	1.05	0.92
Wells Fargo Advantage Intl Equity A	0.48	5.21	1.00	0.90
Average	0.54	5.57	1.09	0.92
S&P Index	0.46	4.67	1.00	1.00

Table 12.9 Performance Data on Bond Funds (2002–2011)

Mutual Fund	Mean Monthly Return	Standard Deviation	Beta	Correlation with Barclays Index
BlackRock Intl Bond Inv A	0.45	2.45	1.31	0.57
Delaware International Bond A	0.79	2.61	1.37	0.56
Federated International Bond A	0.64	2.60	1.47	0.60
Consulting Group International F/I	0.60	1.75	0.98	0.60
Oppenheimer International Bond A	0.87	2.60	1.31	0.54
PACE International Fixed Income A	0.57	2.64	1.49	0.60
T. Rowe Price International Bond	0.65	2.64	1.45	0.59
SEI Instl Intl Tr Intl Fixed-Income A	0.46	1.76	0.90	0.54
Waddell & Reed Global Bond A	0.45	1.38	0.62	0.48
Western Asset Global Government Bond A	0.49	1.43	0.93	0.69
Average	0.60	2.19	1.18	0.58
Barclays US Agg Bond TR USD	0.48	1.07	1.00	1.00

The realized return on international portfolios relative to U.S. portfolios is very dependent on the time period studied. This 10-year period had low returns in the U.S. market. There were other 10-year periods where international portfolios underperformed U.S. portfolios.

There are many fewer international bonds funds than there are stock funds, and their history is much more limited. Table 12.9 shows summary statistics for 11 funds for which 10 years of data are available. The last column is the correlation coefficient of each fund with the β Barclays bond index, which is the standard index used to calculate the performance of U.S. bond funds. It is the bond market equivalent of the S&P index. For U.S. domestic bond funds, the correlation with the Barclays index would be 0.85 to 0.90. Examining the last column shows that the promise of low correlation is met. The average correlation of 0.58 is considerably less than for U.S. bond funds. The standard deviation of a bond fund is very dependent on the maturity of the portfolio. Portfolios of bonds with long maturities have a higher standard of deviation than portfolios of short-maturity bonds. We have no information on the maturity of the foreign bond funds relative to the Barclays index. Thus it is not meaningful to compare standard deviations.

SOVEREIGN FUNDS

Sovereign wealth funds are a relatively new and potentially important institutional development in international finance. Sovereign funds are the investment portfolios of nation-states. The first sovereign funds were created by commodity-rich countries, such as Middle Eastern oil states that transformed natural resources into financial assets. Several non-commodity-dependent nations have also created sovereign funds as a way to manage their currency reserves, or simply as a strategic choice. Some sovereign funds are very large.[19] The following table (from Sovereign Wealth Fund Institute) is a list of the largest sovereign wealth funds by holdings as of 2013. More than half of the largest funds are major oil-producing nations, and the others are countries in the Asia/Pacific region.

[19]Dyck and Morse (2011).

Country	Abbreviation	Fund	Assets $Billion	Inception	Origin
Norway	GPF	Government Pension Fund - Global	664.3	1990	Oil
Abu Dhabi	ADIA	Abu Dhabi Investment Authority	627	1976	Oil
China	SAFE	SAFE Investment Company	567.9**	1997	Non-commodity
Saudi Arabia	SAMA	SAMA Foreign Holdings	532.8	n/a	Oil
China	CIC	China Investment Corporation	482	2007	Non-commodity
Hong Kong	HKMA	Hong Kong Monetary Authority Investment Portfolio	298.7	1993	Non-commodity
Kuwait	KIA	Kuwait Investment Authority	296	1953	Oil
Singapore	GIC	Government of Singapore Investment Corporation	247.5	1981	Non-commodity
Singapore	TH	Temasek Holdings	157.5	1974	Non-commodity
Russia	RNWF	National Welfare Fund	149.7*	2008	Oil
China	NSSF	National Social Security Fund	134.5	2000	Non-commodity
Qatar	QIA	Qatar Investment Authority	115	2003	Oil
Australia	AFF	Future Fund	83	2004	Non-commodity
Dubai	ICD	Investment Corporation of Dubai	70	2006	Oil
Abu Dhabi	IPIC	International Petroleum Investment Company	65.3	1984	Oil
Libya	L1A	Libyan Investment Authority	65	2006	Oil
Kazakhstan	KNF	Kazakhstan National Fund	61.8	2000	Oil
Algeria	RRF	Revenue Regulation Fund	56.7	2000	Oil
Abu Dhabi	MDC	Mubadala Development Company	53.1	2002	Oil
South Korea	KIC	Korea Investment Corporation	43	2005	Non-commodity

Source: Sovereign Wealth Fund Institute (http://www.swfinstitute.org/fund-rankings/)

As of 2013, sovereign wealth funds had nearly $5 trillion in assets, or a little more than 2% of global financial assets. As natural resource extraction continues, this number will likely grow. In the future, sovereign funds may hold a nontrivial fraction of world's wealth. As large investors, these funds will play an increasingly important role in corporate governance and the capital markets.

As entities created and responsible to nation-states, sovereign funds ultimately are the product of a political process. Ang (2010) points out that they derive their purpose and legitimacy from that process.[20] The goals of the fund—and the benchmarks used to assess fund performance—must reflect the fund's purpose. Dyck and Morse (2011) collected data on the

[20]Ang (2010).

world's sovereign funds and studied the way that they invest. They found evidence in favor of Ang's analysis. In aggregate, sovereign fund allocations are partly explained by the nation's strategic industrial plan. Countries with a plan to focus economic development in certain industries, for example, overweigh those industries in their sovereign portfolios as well. This suggests that, at least for some nations, the sovereign fund is one element in a broader strategy for national development. Because of this, their allocation decisions may not always appear optimal from a strictly economic perspective.

The largest of the sovereign funds in 2013 was the Norwegian Pension Fund Global (NPFG) with more than $600 billion of assets. The purpose of the fund is to benefit future generations of Norwegians by retaining some of Norway's revenue from North Sea oil operations in a fund that invests in financial assets. The fund is ultimately controlled by the Norwegian legislature, whose policy is implemented by the Ministry of Finance, which in turn oversees the investment manager, Norges Bank—a branch of the central bank. As of 2013, the legislature specified a 60/40 allocation between global equities and fixed income, reflecting a desire for diversification as well as an emphasis on real returns deriving from an expected future equity risk premium.

The financial crisis of 2008 caused the Norwegian people to revisit the implementation of this policy through an in-depth study of the question of whether the fund should rely on "active" versus "passive" management.[21] Like most financial portfolios, the NPFG lost value during the global crisis. Was active management the culprit? Evidently not. The in-depth report found that most of the loss in value was predictably due to drops in passive indexes of world stocks and bonds—not to poor decisions made by active managers around the crisis.

MODELS FOR MANAGING INTERNATIONAL PORTFOLIOS

For most time periods, empirical evidence suggests that a portfolio of international equities should be a part of an optimum portfolio.

The obvious strategy for an investor deciding to diversify internationally but not wishing to determine how to construct an international portfolio is to hold an international index fund. The parallel to holding a domestic index fund is to hold a value-weighted portfolio of international securities. The Morgan Stanley Capital International index excluding the United States is a value-weighted index, and an investment matching this index would be a value-weighted index fund.

The justification for holding a U.S. index fund rests with the equilibrium models discussed in Chapters 13–16. If expected return is related to a market index and if securities are in equilibrium, then bearing nonmarket or unique risk does not result in additional compensation. The way to eliminate nonmarket risk is to hold an index fund. Even an investor who believes that securities are out of equilibrium but does not profess to know which securities give a positive or negative nonequilibrium return (has no forecasting ability) should hold the index fund. In this case, bearing nonmarket risk, on average, does not improve expected return because the investor, on average, selects securities with zero nonmarket return. Thus the investor should eliminate nonmarket risk by holding an index fund. If there was good evidence that individual securities' expected returns were determined by an international equilibrium model, and if a value-weighted index was the factor affecting expected returns, a parallel argument could be presented for holding an international value-weighted index fund. However, the evidence in favor of any international model determining expected return is still controversial.

[21]Ang, Goetzmann, and Shaefer (2009).

A disturbing aspect of an international index fund is the disproportionate share some countries represent of the world index. If one believes in an international equilibrium asset pricing model, then this is appropriate. Otherwise, it makes sense only if these countries are expected to have an abnormally high return; for diversification or risk arguments it is clearly inappropriate. The authors have heard a number of presentations suggesting other weighting schemes, such as trade or GNP. The correct justification for any weighting should come from equilibrium arguments; otherwise, any weighting is as arbitrary as another.

If one is not willing to accept an international equilibrium model that partitions risk into that part that results in higher expected return and that part that is unique, it is appropriate for an investor without an ability to forecast expected returns to minimize total risk. The risk structure is reasonably predictable through time. The low correlation, on average, among country portfolios, and the pattern of relatively high correlation among countries with close economic links (such as the United States and Canada), is likely to continue in the future. Both Jorion (1985) and Eun and Resnick (1988) have examined the stability of the correlation structure and have found predictability. Thus the past correlation matrices can be used to predict the future. Similarly, Jorion (1985) has shown that standard deviations are predictable, and thus a low-risk international portfolio can be developed.

If one wishes to develop an active international portfolio, then many of the same considerations are involved as are present in developing an active domestic portfolio. However, international investment adds two elements to the investment process not present in pure domestic investment—country selection and exchange exposure.[22]

The decision concerning how much to invest in each country depends on the factors discussed earlier, namely, intercountry correlation, the variance of return for each country's securities, and the expected return in each country. There is good evidence that the past standard deviations and correlations are useful in predicting the future.

Recently a number of researchers have also found some predictability in returns. Harvey (1995), Solnick (1998), and Campbell and Hammo (1992) find predictability in many country's returns. The predictability is low, with 1%–2% of the variation in returns explained by past variables. However, Kandel and Stambaugh (1996) provide evidence that even with this low explanatory power, improvement in portfolio allocation can be achieved. What variables seem to predict returns? Lagged returns, price levels (dividend price, earnings price, and book price ratios), interest rate levels, yield spreads, and default premiums have all been used. How is this done?

In Chapter 8 we discussed how to estimate the coefficients in a multi-index model. For example, we could estimate the relationship between return in a country (e.g., France) and some of the variables that have been found to predict return. Performing this analysis, we could find the relationship

$$\text{Return} = -1 + 1\ (\text{return in the prior period}) + \tfrac{1}{2}\ (\text{interest rate in the prior period})$$

The coefficients, -1, 1, and $\frac{1}{2}$, are estimated by running a time series regression. To forecast return in the next period, one simply substitutes the current value of this period's return and interest rates into the right side of the equation.

[22]Technically, the amount to invest in any security should depend on securities selected in other countries. Thus our treatment of first selecting each portfolio within a country and then doing country selection is nonoptimal. However, it captures much of practice. Furthermore, intercountry factors are relatively unimportant in determining each security's return, so this assumption may be a simplification that improves performance.

These predictions of return plus past values of correlations and standard deviations can be used as input to the portfolio optimization process.

A second possibility for predicting expected returns is to utilize any of the valuation models discussed in Chapter 18. For example, the infinite constant growth model states that

$$\text{Expected return} = \frac{\text{Dividend}}{\text{Price}} + \text{Growth}$$

Estimates of next period's dividend could be obtained by estimating earnings and estimating the proportion of earnings paid out as dividends (the payout rate). The payout ratio for a country portfolio is very stable over time, and forecasts of earnings are widely available and at an economy level quite accurate. Estimates of growth rates in earnings are also widely available internationally. Thus valuation models are a feasible way to estimate expected returns.[23]

One of the few studies that examines some alternative ways of estimating expected return is Arnott and Henriksson (1989). They forecast the relative performance of each country's stocks compared to the country's bonds on the basis of current risk premiums and economic variables. They define the risk premium as the difference in expected return between common equity and bonds. They measure expected return on bonds by using the yield to maturity. They measure expected return on equity by calculating the earnings divided by price. Comparing this measure with the valuation model just presented shows that growth should be added and differences in payout taken into account. These differences, as well as differences in accounting conventions across countries and the impact of this on earnings, could affect risk premium comparisons across countries. They recognize these influences, and instead of using risk premiums directly, they use current risk premiums relative to past risk premiums. Their forecast equation states that future performance is related to current risk premiums divided by average risk premiums in the past. In equation form this is

$$\text{Future returns on equities relative to debt} = \text{Constant}$$

$$+ \text{Constant (Current risk premium/Average risk premium prior two years)}$$

They find for many countries that this equation is a useful predictor and that for some countries it can be improved by adding other macroeconomic variables, such as prediction of trade and production statistics. This model could be used to estimate which countries have higher expected future returns on equities by using current bond yields as expected returns for bonds and the preceding equation to estimate the difference between bond and equity returns. Clearly further testing of all of these models is necessary. However, they are suggestive of the type of analysis that can be done in active international asset allocation.

The second new consideration that international investment introduces is exchange risk. As discussed earlier, entering into futures contracts can reduce the variability because of the exchange risk. Considering only risk, this is generally useful. Entering into futures contracts can also affect expected return, however. As discussed in Chapter 24, entering into a futures contract could lower expected returns. Furthermore, the investor may have some beliefs about changes in exchange rates different from those contained in market prices.[24] In this case the sacrifice in expected return may lead the investor to choose not to eliminate exchange risk.

---

[23]Testing of the accuracy of forecasts produced by these models is unavailable, so all we can do is to suggest types of analysis; we cannot report results.

[24]Levich (1970, 1979) has shown that some forecasters are able to predict exchange rate movements.

Finally, Black (1989) has shown that taking some exchange risk can increase expected return. Thus exchange rate exposure involves a risk–return trade-off.

Active Short-Term Bond Management

Risk-free interest rates differ from country to country. For example, the interest rate on six-month government issues could be 7% in England and 4% in the United States. The expected return for a U.S. investor buying an English bond would be the expected return to a British investor plus the exchange gains and losses.

Theory says the exchange gain or loss should be related to the interest rate differential. Thus the U.S. investor should expect to lose about 3% in exchange rate changes by buying the British bond. However, empirical evidence does not support the claim that exchange rate changes have a close relationship to interest rate differentials.

The empirical evidence strongly supports that investment in the high–interest rate country gives the higher return.[25] Three explanations have been suggested: a peso explanation, extra risk, and an investment opportunity. The peso explanation is named after the investors who invested their money in Mexican government bonds. For a number of years they earned a return greater than they would have earned in the United States. When the devaluation occurred, however, it more than eliminated all past gains. The peso argument is that although the empirical evidence suggests gains by investing in the higher–interest rate countries, some future devaluation will eliminate all gains. The return gains have been so persistent that the size of a devaluation necessary to eliminate past gains seems too large to be plausible. Thus most analysts reject this explanation.

The second explanation is that the extra return is simply compensation for risk. Although some of the extra return may be compensation for risk, studies to date do not support this as a complete explanation. Thus there seems to be an investment opportunity, and there are a number of funds that follow the strategy of investing in the higher-yielding country.[26]

CONCLUSION

In this chapter we have discussed the evidence in support of international diversification. The evidence that international diversification reduces risk is uniform and extensive. Given the low risk, international diversification is justified even if expected returns are less internationally than domestically. Unless there are mechanisms, such as taxes or currency restrictions, that substantially reduce the return on foreign investment relative to domestic investment, international diversification has to be profitable for investors of some countries, and possibly all.

---

[25]For example, Cumby and Glen (1990) find on average that exchange rate changes increase the return of buying the higher–interest rate countries (e.g., British bonds) would be expected to return more than 7%.

[26]There is a variation in this strategy that some funds follow. Assume we observe the following interest rates on six-month government debt:

U.S. rate = 4%

English rate = 7%

German rate = 5%

In this scenario, one investment strategy is to buy English bonds and hedge exchange risk by buying a futures contract of euros for dollars. The investor will lose 1% on the futures contract since there is a 1% difference in T-bill rates and empirical evidence supports that the interest rate differential is reflected in the futures contract. If the English bond–euro exchange rate stays constant, the investor will earn 7% on the bond less 1% on the futures contract, or 6%, which is superior to the return on U.S. bills.

QUESTIONS AND PROBLEMS

1. Assume that you expect that the average return on a security in various markets is as shown in the following table. Assume further that the historical correlation coefficients shown in Table 12.1 are a reasonable estimate of future correlation coefficients. Finally, assume the standard deviations shown in Table 12.2. Which markets are attractive investments for an American investor if the riskless lending and borrowing rate is 6%?

	Market	Expected Return (%)
1.	Australia	14
2.	France	16
3.	Japan	14
4.	United Kingdom	15
5.	United States	15

2. Consider the following returns:

Period	United States	United Kingdom	Exchange Rate[a]
1	10%	5%	$3
2	15%	−5%	2.5
3	−5%	15%	2.5
4	12%	8%	2.0
5	6%	10%	1.5
6			2.5

[a]Beginning of period dollars for pounds.

What is the average return in each market from the point of view of a U.S. investor and of a U.K. investor?

3. Given the data in the prior question, what is the standard deviation of return from the point of view of a U.S. investor and of a U.K. investor?

4. For the following returns:

Period	United States	Japan	Exchange Rate[a]
1	12%	18%	200
2	15%	12%	180
3	5%	10%	190
4	10%	12%	150
5	6%	7%	170
6			180

[a]Beginning of period value of yen for dollars.

What is the average return in each market from the point of view of a U.S. and a Japanese investor?

5. What is the standard deviation of return from the point of view of a U.S. and a Japanese investor?

6. What is the correlation of return between markets from the point of view of each investor?

BIBLIOGRAPHY

1. Adler, Michael. "The Cost of Capital and Valuation of a Two-Country Firm," *Journal of Finance*, **XXIX,** No. 1 (March 1974), pp. 119–132.
2. Adler, Michael, and Horesh, Reuven. "The Relationship among Equity Markets: Comment on [3]," *Journal of Finance*, **XXIX,** No. 4 (Sept. 1974), pp. 1131–1317.
3. Adler, Michael, and Dumas, Bernard. "International Portfolio Choice and Corporate Finance: A Synthesis," *Journal of Finance*, **38,** No. 3 (June 1983), pp. 925–984.
4. Adler, Michael, and Prasad, Bhaskar. "On Universal Currency Hedges," *Journal of Financial and Quantitative Analysis*, **27,** No. 1 (March 1992), pp. 19–38.
5. Agmon, Tamir. "The Relations among Equity Markets: A Study of Share Price Co-movements in the United States, United Kingdom, Germany and Japan," *Journal of Finance*, **XXVII,** No. 3 (June 1972), pp. 839–855.
6. ——. "Country Risk: The Significance of the Country Factor for Share-Price Movements in the United Kingdom, Germany, and Japan," *Journal of Business*, **46,** No. 1 (Jan. 1973), pp. 24–32.
7. ——. "Reply to [2]," *Journal of Finance*, **XXIX,** No. 4 (Sept. 1974), pp. 1318–1319.
8. Ang, Andrew. "The Four Benchmarks of Sovereign Wealth Funds," (Sept. 21, 2010).
9. Ang, A., Goetzmann, W., and Shaefer, S., "Evaluation of Active Management of the Norwegian Government Pension Fund–Global," Norwegian Ministry of Finance (2009), http://www.regjeringen.no/upload/FIN/Statens%20pensjonsfond/rapporter/AGS%20Report.pdf
10. Agmon, Tamir, and Lessard, Donald. "Investor Recognition of Corporate International Diversification," *Journal of Finance*, **XXXII,** No. 4 (Sept. 1977), pp. 1049–1055.
11. Arnott, A., and Henriksson, N. "A Disciplined Approach to Global Asset Allocation," *Financial Analyst Journal* (March–April 1989), pp. 17–28.
12. Baxter, Marianne. "The International Diversification Puzzle Is Worse Than You Think," *The American Economic Review*, **87,** No. 1 (March 1997), pp. 170–180.
13. Bekaert, G., and Harvey, C. "Time-Varying World Market Integration." *Journal of Finance*, **50** (1995), pp. 403–44.
14. Bekaert, Geert, and Campbell, R. Harvey. "Foreign Speculators and Emerging Equity Markets." *Journal of Finance*, **55**, No. 2 (2000), pp. 565–613.
15. Bennett, James A. "International Stock Market Equilibrium with Heterogenous Tastes," *The American Economic Review*, **89,** No. 3 (June 1999), pp. 639–648.
16. Black, F. "International Capital Market Equilibrium with Investment Barriers," *Journal of Financial Economics*, **1,** No. 4 (Dec. 1974), pp. 337–352.
17. Black, F. "Equilibrium Exchange Rate Hedging," National Bureau of Economic Research (NBER) working paper, No. 2947 (April 1989).
18. Branch, Ben. "Common Stock Performance and Inflation: An International Comparison," *Journal of Business*, **47,** No. 1 (Jan. 1973), pp. 48–52.
19. Campbell, J., and Hammo, Y. "Predictable Stock Returns in the United States and Japan: A Study of Long-Term Capital Market Integration," *Journal of Finance*, **47** (1992), pp. 43–70.
20. Chen, Zhiwu, and Knez, Peter J. "Measurement of Market Integration and Arbitrage," *Review of Financial Studies*, **8,** No. 2 (1995), pp. 287–325.
21. Cho, Chinhyung D., Eun, Cheol S., and Senbet, Lemma. "International Arbitrage Pricing Theory: An Empirical Investigation," *Journal of Finance*, **41,** No. 2 (June 1986), pp. 313–329.
22. Chollerton, Kenneth, Pieraerts, Pierre, and Solnik, Bruno. "Why Invest in Foreign Currency Bonds?" *Journal of Portfolio Management*, **22** (Summer 1986), pp. 4–8.
23. Cumby, Robert. "Is It Risk? Explaining Deviations from Uncovered Interest Rate Parity," *Journal of Monetary Economics*, **22,** No. 2 (1988), pp. 297–300.
24. Cumby, Robert, and Glen, Jack. "Evaluating the Performance of International Mutual Funds," *Journal of Finance*, **24** (1990), pp. 408–435.
25. Dyck, I. J. Alexander, and Morse, Adair. "Sovereign Wealth Fund Portfolios," Chicago Booth Research Paper No. 11–15; MFI working paper No. 2011–003; Rotman School of Management working paper No. 1792850 (Feb. 1, 2011).
26. Elton, Edwin J., Gruber, Martin J., and Rentzler, Joel. "Professionally Managed, Publicly Traded Community Funds," *The Journal of Business*, **60,** No. 2 (April 1987), pp. 175–199.

27. Eun, Cheol, Kolodny, Richard, and Resnick, Bruce. "U.S. Based International Mutual Funds: A Performance Evaluation," *Journal of Portfolio Management* (Spring 1991), pp. 88–94.
28. Eun, Cheol S., and Resnick, Bruce G. "Exchange Rate Uncertainty, Forward Contracts, and International Portfolio Selection," *The Journal of Finance*, **43,** No. 1 (March 1988), pp. 197–215.
29. Fama, Eugene, and French, Kenneth. "Business Conditions and Expected Return on Stocks and Bonds," *Journal of Financial Economics*, **25** (1993), pp. 23–50.
30. Farber, Andre L. "Performance of Internationally Diversified Mutual Funds," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
31. Fatemi, Ali M. "Shareholder Benefits from Corporate International Diversification," *The Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1325–1344.
32. French, Kenneth, and Poterba, James. "Japanese and U.S. Cross-Border Common Stock Investments," *Journal of the Japanese and International Economics*, **4** (Dec. 1990), pp. 476–493.
33. French, Kenneth R., and Poterba, James M. "Investor Diversification and International Equity Markets," *The American Economic Review*, **81,** No. 2 (May 1991), pp. 222–226.
34. Froot, Kenneth A., and Dabora, Emil M. "How Are Stock Prices Affected by the Location of Trade?" *Journal of Financial Economics*, **53,** No. 2 (Aug. 1999), pp. 189–216.
35. Goetzmann, William N., and Jorion, Philippe. "Re-emerging Markets," *The Journal of Financial and Quantitative Analysis*, **34,** No. 1 (March 1999), pp. 1–32.
36. Goetzmann, William N., Li, Lingfeng, and Rouwenhorst, K. Geert. "Long-Term Global Market Correlations," *The Journal of Business*, **78,** No. 1 (Jan. 2005), pp. 1–38.
37. Goetzmann, William N., and Ukhov, Andrey. "British Investment Overseas 1870–1913: A Modern Portfolio Theory Approach," *Review of Finance*, **10,** No. 2 (2006), pp. 261–300.
38. Grauer, R., and Hakansson, Nils. "Gains from Internation Diversification: 1968–85 Returns on Portfolios of Stocks and Bonds," *Journal of Finance* (July 1987), pp. 721–738.
39. Grauer, Robert R., Hakansson, Nils H., and Crouhy, Michel. "Gains from International Diversification: 1968–85 Returns on Portfolios of Stocks and Bonds/Discussion," *The Journal of Finance*, **42,** No. 3 (July 1987), pp. 721–741.
40. Grauer, F., Litzenberger, R., and Stehle, R. "Sharing Rules and Equilibrium in an International Capital Market under Uncertainty," *Journal of Financial Economics*, **3,** No. 3 (June 1976), pp. 233–256.
41. Grubel, Herbert. "Internally Diversified Portfolios: Welfare Gains and Capital Flows," *American Economic Review*, **LVIII,** No. 5, Part 1 (Dec. 1968), pp. 1299–1314.
42. Grubel, G. Herbert, and Fadner, Kenneth. "The Interdependence of International Equity Markets," *Journal of Finance*, **XXVI,** No. 1 (March 1971), pp. 89–94.
43. Gultekin, N. Bulent. "Stock Market Returns and Inflation: Evidence from Other Countries," *The Journal of Finance*, **38,** No. 1 (March 1983), pp. 49–68.
44. Guy, J. "The Performance of the British Investment Trust Industry," *Journal of Finance*, **33** (May 1978), pp. 443–455.
45. Harvey, Campbell R. "Predictable Risk and Returns in Emerging Markets," *Review of Financial Studies*, **8,** No. 3 (1995), pp. 773–816.
46. Heston, Steven L., and Rouwenhorst, K. Geert. "Does Industrial Structure Explain the Benefits of International Diversification?" *Journal of Financial Economics*, **36,** No. 1 (Aug. 1994), pp. 3–27.
47. Huberman, Gur. "Familiarity Breeds Investment," *Review of Financial Studies*, **14,** No. 3 (2001), pp. 659–680.
48. Ibbotson, Roger, Siegal, Lawrence, and Love, Kathryn. "World Wealth: Market Values and Returns," *Journal of Portfolio Management*, **4,** No. 2 (Fall 1985), pp. 4–23.
49. Jacquillat, Bertrand, and Solnik, Bruno. "Multi-Nationals Are Poor Tools for Diversification," *Journal of Portfolio Management*, **11,** No. 1 (Winter 1978), pp. 8–12.
50. Jorion, Philippe. "International Diversification with Estimation Risk," *Journal of Business*, **12,** No. 1 (July 1985), pp. 259–278.
51. Joy, Maurice, Panton, Don, Reilly, Frank, and Martin, Stanley. "Co-Movements of International Equity Markets," *The Financial Review*, **58,** No. 3 (1976), pp. 1–20.

52. Kandel, Shmuel, and Stambaugh, Robert. "On the Predictability of Stock Returns: An Asset-Allocation Perspective," *Journal of Finance*, **51** (1996) pp. 385–424.
53. Kaplanis, C. E., and Schaefer, Steve. "Exchange Risk and International Diversification in Bond and Equity Portfolios," *Journal of Economics and Business*, **43,** No. 4 (1991), pp. 287–307.
54. Lessard, Donald. "International Portfolio Diversification: A Multivariate Analysis for a Group of Latin American Countries," *Journal of Finance*, **XXVIII,** No. 3 (June 1973), pp. 619–633.
55. ——. "World, National and Industry Factors in Equity Returns," *Journal of Finance*, **XXIV,** No. 2 (May 1974), pp. 379–391.
56. ——. "The Structure of Returns and Gains from International Diversification: A Multivariate Approach," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
57. Levich, Richard. "On the Efficiency of Markets for Foreign Exchange," in J. Frenkel and R. Dornbusch (eds.), *International Economic Policy: Theory and Evidence*, Vol. 42 (Baltimore, MD: Johns Hopkins University Press, 1970).
58. ——. "The Efficiency of Markets for Foreign Exchange: A Review and Extension," in Donald Lessard (ed.), *International Financial Management: Theory and Application* (New York: Warren, Gorham, and Lamont, 1979).
59. Levich, Richard, and Frenkel, Jacob. "Covered Interest Arbitrage: Unexplored Profits?" *Journal of Political Economy* (April 1975), pp. 325–338.
60. ——. "Transaction Costs and Interest Arbitrage: Tranquil versus Turbulent Periods," *Journal of Political Economy* (Dec. 1977), pp. 1209–1286.
61. Levy, Haim, and Sarnat, Marshall. "International Diversification of Investment Portfolios," *American Economic Review*, **LX,** No. 4 (Sept. 1970), pp. 668–675.
62. ——. "Devaluation Risk and the Portfolio Analysis of International Investment," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
63. Lowenfeld, H. "Investment an Exact Science," *The Financial Review of Reviews*.
64. Makin, John. "Portfolio Theory and the Problem of Foreign Exchange Risk," *Journal of Finance*, **XXXIII,** No. 2 (May 1978), pp. 517–534.
65. McDonald, John. "French Mutual Fund Performance: Evaluation of Internationally-Diversified Portfolios," *Journal of Finance*, **XXVIII,** No. 5 (Dec. 1973), pp. 1161–1180.
66. Obstfeld, Maurice. "Risk-Taking, Global Diversification, and Growth," *The American Economic Review*, **84,** No. 5 (Dec. 1994), pp. 1310–1329.
67. Panton, Don, Lessig, Parket, and Joy, Maurice. "Co-movement of International Equity Markets: A Taxonomic Approach," *Journal of Financial and Quantitative Analysis*, **XI,** No. 3 (Sept. 1976), pp. 415–432.
68. Quinn, Dennis P., and Voth, Hans-Joachim. "A Century of Global Equity Market Correlations," *The American Economic Review*, **98,** No. 2 (May 2008), pp. 535–540.
69. Rosenthal, L., and Young, C. "The Seemingly Anomalous Price Behavior of Royal Dutch Shell and Unilever N. V. PLC," *Journal of Financial Economics*, **26** (1990), pp. 123–141.
70. Roxbugh, Charles, Lund, Susan, and Piotrowski, John. "Mapping Global Capital Markets 2011," McKinsey Global Institute (2011).
71. Ripley, Duncan. "Systematic Elements in the Linkage of National Stock Market Indices," *Review of Economics and Statistics*, **LV,** No. 3 (Aug. 1973), pp. 356–361.
72. Robichek, Alexander, and Eaker, Mark. "Foreign Exchange Hedging and the Capital Asset Pricing Model," *Journal of Finance*, **XXXIII,** No. 3 (June 1978), pp. 1011–1018.
73. Severn, Alan. "Investor Evaluation of Foreign and Domestic Risk," *Journal of Finance*, **XXIX,** No. 2 (May 1974), pp. 545–550.
74. Sharma, J. L., and Kennedy, Robert. "A Comparative Analysis of Stock Price Behavior on the Bombay, London, and New York Stock Exchanges," *Journal of Financial and Quantitative Analysis*, **XII,** No. 3 (Sept. 1977), pp. 391–413.
75. Solnik, Bruno. "The International Pricing of Risk: An Empirical Investigation of the World Capital Market Structure," *Journal of Finance*, **XXIX,** No. 2 (May 1974), pp. 365–378.

76. ——. "Why Not Diversify Internationally?" *Financial Analysts Journal*, **20,** No. 4 (July/Aug. 1974), pp. 48–54.
77. ——. "An Equilibrium Model of the International Capital Market," *Journal of Economic Theory*, **8,** No. 4 (Aug. 1974), pp. 500–524.
78. ——. "An International Market Model of Security Price Behavior," *Journal of Financial and Quantitative Analysis*, **IX,** No. 4 (Sept. 1974), pp. 537–554.
79. ——. "The Advantages of Domestic and International Diversification," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
80. ——. "Testing International Asset Pricing: Some Pessimistic Views," *Journal of Finance*, **XXXII,** No. 2 (May 1977), pp. 503–512.
81. Solnik, Bruno. *International Investments* (Reading, MA: Addison-Wesley, 1988).
82. ——. "The Performance of International Asset Allocations Strategies Using Conditioning Information," *Journal of Empirical Finance*, **1,** No. 1 (June 1993), pp. 33–55.
83. ——. "Global Asset Management," *The Journal of Portfolio Management* (Summer 1998), pp. 43–51.
84. Solnik, Bruno, and Noetzlin, B. "Optimal International Asset Allocation," *Journal of Portfolio Management*, **2** (Fall 1982), pp. 11–21.
85. Stehle, Richard. "An Empirical Test of the Alternative Hypotheses of National and International Pricing of Risky Assets," *Journal of Finance*, **XII,** No. 2 (May 1977), pp. 493–502.
86. Subrahmanyam, Marti. "International Capital Markets, Equilibrium, and Investor Welfare with Unequal Interest Rates," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
87. ——. "On the Optimality of International Capital Market Integration," *Journal of Financial Economics*, **2,** No. 1 (March 1975), pp. 3–28.
88. Uppal, Raman. "A General Equilibrium Model of International Portfolio Choice," *The Journal of Finance*, **48,** No. 2 (June 1993), pp. 529–553.

Part 3

MODELS OF EQUILIBRIUM IN THE CAPITAL MARKETS

13

The Standard Capital Asset Pricing Model

All of the preceding chapters have been concerned with how an individual or institution, acting on a set of estimates, could select an optimum portfolio, or set of portfolios. If investors act as we have prescribed, then we should be able to draw on the analysis to determine how the aggregate of investors will behave and how prices and returns at which markets will clear are set. The construction of general equilibrium models will allow us to determine the relevant measure of risk for any asset and the relationship between expected return and risk for any asset when markets are in equilibrium. Furthermore, though the equilibrium models are derived from models of how portfolios should be constructed, the models themselves have major implications for the characteristics of optimum portfolios.

The subject of equilibrium models is so important that we have devoted four chapters to it. In this chapter we develop the simplest form of an equilibrium model, called the *standard capital asset pricing model* (CAPM), or the *one-factor capital asset pricing model*. This was the first general equilibrium model developed, and it is based on the most stringent set of assumptions. The next chapter, on general equilibrium models, deals with models that have been developed under more realistic sets of assumptions. The third chapter in this sequence deals with tests of general equilibrium models. The final chapter deals with a new theory of asset pricing: arbitrage pricing theory.

It is worthwhile pointing out at this time that the final test of a model is not how reasonable the assumptions behind it appear but how well the model describes reality. As readers proceed with this chapter, they will, no doubt, find many of its assumptions objectionable. Furthermore, the final model is so simple that readers may well wonder about its validity. As we will see, despite the stringent assumptions and the simplicity of the model, it does an amazingly good job of describing prices in the capital markets.

THE ASSUMPTIONS UNDERLYING THE STANDARD CAPITAL ASSET PRICING MODEL (CAPM)

The real world is sufficiently complex that to understand it and construct models of how it works, one must assume away those complexities that are thought to have only a small (or no) effect on its behavior. As the physicist builds models of the movement of matter in a frictionless environment, the economist builds models where there are no institutional frictions to the movement of stock prices.

The first assumption we make is that there are no transaction costs. There is no cost (friction) of buying or selling any asset. If transaction costs were present, the return from any asset would be a function of whether the investor owned it before the decision period. Thus to include transaction costs in the model adds a great deal of complexity. Whether it is worthwhile introducing this complexity depends on the importance of transaction costs to investors' decisions. Given the size of transaction costs, they are probably of minor importance.

The second assumption behind the CAPM is that assets are infinitely divisible. This means that investors could take any position in an investment, regardless of the size of their wealth. For example, they can buy one dollar's worth of IBM stock.

The third assumption is the absence of personal income tax.[1] This means, for example, that the individual is indifferent to the form (dividends or capital gains) in which the return on the investment is received.

The fourth assumption is that an individual cannot affect the price of a stock by his buying or selling action. This is analogous to the assumption of perfect competition. Although no single investor can affect prices by an individual action, investors in total determine prices by their actions.

The fifth assumption is that investors are expected to make decisions solely in terms of expected values and standard deviations of the returns on their portfolios. In other words, they make their portfolio decision utilizing the framework discussed in other chapters.

The sixth assumption is that unlimited short sales are allowed. The individual investor can sell short any number of any shares.[2]

The seventh assumption is unlimited lending and borrowing at the riskless rate. The investor can lend or borrow any amount of funds desired at a rate of interest equal to the rate for riskless securities.

The eighth and ninth assumptions deal with the homogeneity of expectations. First, investors are assumed to be concerned with the mean and variance of returns (or prices over a single period), and all investors are assumed to define the relevant period in exactly the same manner. Second, all investors are assumed to have identical expectations with respect to the necessary inputs to the portfolio decision. As we have said many times, these inputs are expected returns, the variance of returns, and the correlation matrix representing the correlation structure between all pairs of stocks.

The tenth assumption is that all assets are marketable. All assets, including human capital, can be sold and bought on the market.

Readers can now see the reason for the earlier warning that they might find many of the assumptions behind the CAPM untenable. It is clear that these assumptions do not hold in the real world, just as it is clear that the physicist's frictionless environment does not really exist. The relevant questions are, How much is reality distorted by making these assumptions? What conclusions about capital markets do they lead to? Do these conclusions seem to describe the actual performance of the capital market?

THE CAPM

The standard form of the general equilibrium relationship for asset returns was developed independently by Sharpe, Lintner, and Mossin. Hence it is often referred to as the *Sharpe–Lintner–Mossin form* of the capital asset pricing model. This model has been

[1]The major results of the model would hold if income tax and capital gains taxes were of equal size.

[2]This model can be derived under either of the descriptions of short sales discussed in Chapter 5.

derived in several forms involving different degrees of rigor and mathematical complexity. There is a trade-off between these derivations. The more complex forms are more rigorous and provide a framework within which alternative sets of assumptions can be examined. However, because of their complexity, they do not convey the economic intuition behind the CAPM as readily as some of the simpler forms. Because of this, we approach the derivation of the model at two distinct levels. The first derivation consists of a simple, intuitively appealing derivation of the CAPM. This is followed by a more rigorous derivation.

Deriving the CAPM—A Simple Approach

Recall that in the presence of short sales, but without riskless lending and borrowing, each investor faced an efficient frontier such as that shown in Figure 13.1. In this figure, *BC* represents the efficient frontier, while *ABC* represents the set of minimum-variance portfolios. In general the efficient frontier will differ among investors because of differences in expectations.

When we introduced riskless lending and borrowing, we showed that the portfolio of risky assets that any investor would hold could be identified without regard to the investor's risk preferences. This portfolio lies at the tangency point between the original efficient frontier of risky assets and a ray passing through the riskless return (on the vertical axis). This is depicted in Figure 13.2, where P_i denotes investor i's portfolio of risky assets.[3] The investors satisfy their risk preferences by combining portfolio P_i with lending or borrowing.

If all investors have homogeneous expectations and they all face the same lending and borrowing rate, then they will each face a diagram such as in Figure 13.2 and, furthermore, all of the diagrams will be identical. The portfolio of risky assets P_i held by any investor will be identical to the portfolio of risky assets held by any other investor. If all investors

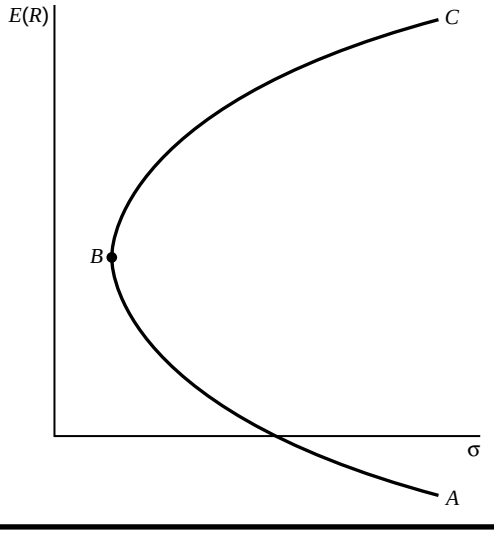


Figure 13.1 The efficient frontier—no lending and borrowing.

[3]We have subscripted *P* because each individual can face a different efficient frontier and thus select a different P_i. This is true, though the composition of P_i does not depend on investor i's risk preference.

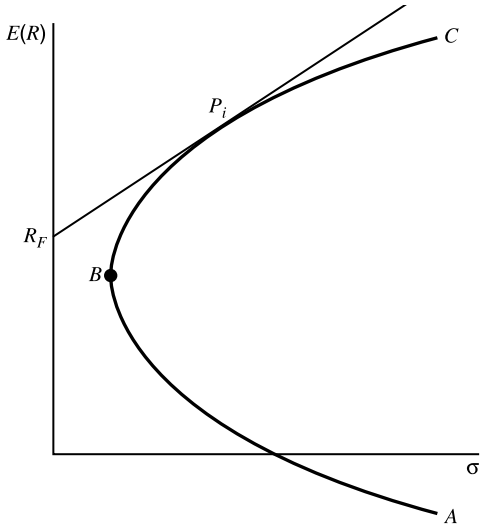


Figure 13.2 The efficient frontier with lending and borrowing.

hold the same risky portfolio, then, in equilibrium, it must be the market portfolio. The market portfolio is a portfolio comprising all risky assets. Each asset is held in the proportion that the market value of that asset represents of the total market value of all risky assets. For example, if IBM stock represents 3% of all risky assets, then the market portfolio contains 3% IBM stock, and each investor will take 3% of the money that will be invested in all risky assets and place it in IBM stock.

Notice that we have already learned something important. All investors will hold combinations of only two portfolios: the market portfolio (M) and a riskless security. This is sometimes referred to as the *two mutual fund theorem* because all investors would be satisfied with a market fund, plus the ability to lend or borrow a riskless security.

The straight line depicted in Figure 13.2 is usually referred to as the *capital market line*. All investors will end up with portfolios somewhere along the capital market line, and all *efficient portfolios* would lie along the capital market line. However, not all securities or portfolios lie along the capital market line. In fact, from the derivation of the efficient frontier, we know that all portfolios of risky and riskless assets, except those that are efficient, lie below the capital market line. By looking at the capital market line, we can learn something about the market price of risk. In Chapter 5 we showed that the equation of a line connecting a riskless asset and a risky portfolio (the line we now call the capital market line) is

$$\bar{R}_e = R_F + \frac{\bar{R}_M - R_F}{\sigma_M}\sigma_e$$

where the subscript e denotes an efficient portfolio.

The term $[(\bar{R}_M - R_F)/\sigma_M]$ can be thought of as the market price of risk for all efficient portfolios.[4] It is the extra return that can be gained by increasing the level of risk (standard deviation) on an efficient portfolio by one unit. The second term on the right-hand side of

[4]The reader should be alerted to the fact that many authors have defined $(\bar{R}_M - R_F)/\sigma_M^2$ as the market price of risk. The reason we have selected $(\bar{R}_M - R_F)/\sigma_M$ will become clear as you proceed with this chapter.

this equation is simply the market price of risk times the amount of risk in a portfolio. The second term represents that element of required return that is due to risk. The first term is simply the price of time or the return that is required for delaying potential consumption, one period given perfect certainty about the future cash flow. Thus the expected return on an efficient portfolio is

$$\text{(Expected return)} = \text{(Price of time)} + \text{(Price of risk)} \times \text{(Amount of risk)}$$

Although this equation establishes the return on an efficient portfolio, it does not describe equilibrium returns on nonefficient portfolios or on individual securities. We now turn to the development of a relationship that does so.

In Chapter 7 we argued that, for very well-diversified portfolios, beta was the correct measure of a security's risk. For *very* well-diversified portfolios, nonsystematic risk tends to go to zero, and the only relevant risk is systematic risk measured by beta. As we have just explained, given the assumptions of homogeneous expectations and unlimited riskless lending and borrowing, all investors will hold the market portfolio. Thus the investor will hold a *very* well-diversified portfolio. Because we assume that the investor is concerned only with expected return and risk, the only dimensions of a security that need be of concern are expected return and beta.

Let us hypothesize two portfolios with the characteristics shown here:

Investment	Expected Return	Beta
A	10	1.0
B	12	1.4

We have already seen (Chapter 5) that the expected return from portfolio *A* is simply the sum of the products of the proportion invested in each stock and the expected return on each stock. We have also seen that the beta on a portfolio is simply the sum of the product of the proportion invested in each stock times the beta on each stock. Now consider a portfolio *C* made up of one-half of portfolio *A* and one-half of portfolio *B.* From the facts stated earlier, the expected return on this portfolio is 11, and its beta is 1.2. These three potential investments are plotted in Figure 13.3. Notice they lie on a straight line. This is no accident. All portfolios composed of different fractions of investments *A* and *B* will lie along a straight line in expected return beta space.[5]

Now hypothesize a new investment *D* that has a return of 13% and a beta of 1.2. Such an investment cannot exist for very long. All decisions are made in terms of risk and return. This portfolio offers a higher return and the same risk as portfolio *C.* Hence it would pay all investors to sell *C* short and buy *D.* Similarly, if a security were to exist with a return

[5]If we let X stand for the fraction of funds invested in portfolio *A*, then the equation for return is

$$\bar{R}_P = X\bar{R}_A + (1 - X)\bar{R}_B$$

The equation for beta is

$$\beta_P = X\beta_A + (1 - X)\beta_B$$

Solving the second equation for X and substituting in the first equation, we see that we are left with an equation of the form

$$\bar{R}_P = a + b\beta_P$$

or the equation of a straight line.

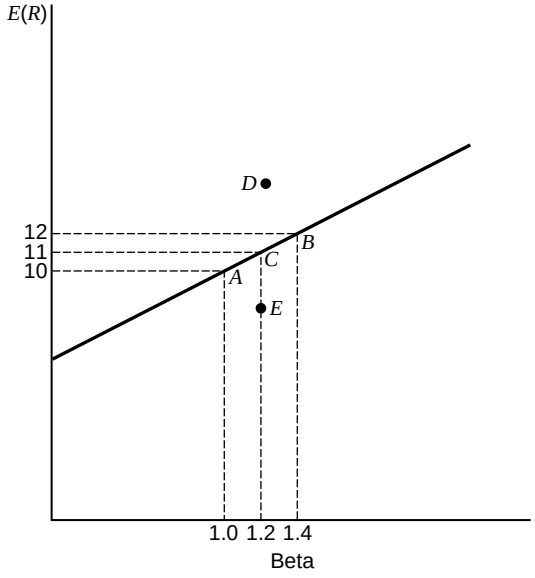


Figure 13.3 Combinations of portfolios.

of 8% and a beta of 1.2 (designated by *E*), it would pay arbitragers to step in and buy portfolio *C* while selling security *E* short. Such arbitrage would take place until *C*, *D*, and *E* all yielded the same return. This is just another illustration of the adage that two things that are equivalent cannot sell at different prices. We can demonstrate the arbitrage discussed earlier in a slightly more formal manner. Let us return to the arbitrage between portfolios *C* and *D*. An investor could sell \$100 worth of portfolio *C* short and with the \$100 buy portfolio *D*. If the investor were to do so, the characteristics of this arbitraged portfolio would be as follows:

	Cash Invested	Expected Return	Beta
Portfolio *C*	−\$100	−\$11	−1.2
Security *D*	+\$100	\$13	1.2
Arbitrage portfolio	0	\$ 2	0

From this example it is clear that as long as a security lies above the straight line, there is a portfolio involving zero risk and zero net investment that has a positive expected profit. An investor will engage in this arbitrage as long as any security or portfolio lies above the straight line depicted in Figure 13.3. A similar arbitrage will exist if any amount lies below the straight line in Figure 13.3.

We have now established that all investments and all portfolios of investments must lie along a straight line in return-beta space. If any investment were to lie above or below that straight line, then an opportunity would exist for riskless arbitrage. This arbitrage would continue until all investments converged to the line. There are many different ways that this straight line can be identified, for it takes only two points to identify a straight line. Because we have shown that, under the assumptions of the CAPM, everybody will hold the market portfolio because all portfolios must lie on the straight line, we will use this as one point. Recall in Chapter 7 we showed that the market portfolio must have a beta of 1.

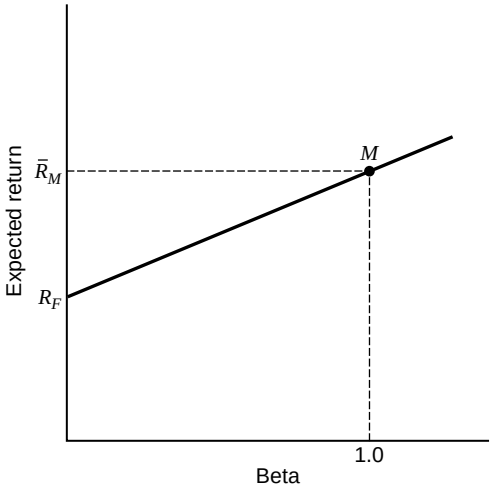


Figure 13.4 The security market line.

Thus, in Figure 13.4, the market portfolio is point *M* with a beta of 1 and an expected return of $\bar{R}_M$. It is often convenient to choose the second point to identify a straight line as the intercept. The intercept occurs when beta equals zero, or when the asset has zero systematic risk. One asset with zero systematic risk is the riskless asset. Thus we can treat the intercept as the rate of return on a riskless asset. These two points identify the straight line shown in Figure 13.4. The equation of a straight line has the form

$$\bar{R}_i = a + b\beta_i \tag{13.1}$$

One point on the line is the riskless asset with a beta of zero. Thus

$$R_F = a + b(0)$$

or

$$R_F = a$$

A second point on the line is the market portfolio with a beta of 1. Thus

$$\bar{R}_M = a + b(1)$$

or

$$\left(\bar{R}_M - a\right) = b$$

Putting these together and substituting into Equation (13.1) yields

$$\bar{R}_i = R_F + \beta_i\left(\bar{R}_M - R_F\right) \tag{13.2}$$

Think about this relationship for a moment. It represents one of the most important discoveries in the field of finance. Here is a simple equation, called the *security market line*, that describes the expected return for all assets and portfolios of assets in the economy. The expected return on any asset, or portfolio, whether it is efficient or not, can be determined

from this relationship. Notice that $\bar{R}_M$ and R_F are not functions of the assets we examine. Thus the relationship between the expected return on any two assets can be related simply to their difference in beta. The higher beta is for any security, the higher must be its equilibrium return. Furthermore, the relationship between beta and expected return is linear. One of the greatest insights that comes from this equation arises from what it states is unimportant in determining return. Recall that in Chapter 7 we saw that the risk of any stock could be divided into systematic and unsystematic risk. Beta was the index of systematic risk. This equation validates the conclusion that systematic risk is the only important ingredient in determining expected returns and that nonsystematic risk plays no role.[6] Put another way, the investor gets rewarded for bearing systematic risk. It is not total variance of returns that affects expected returns but only that part of the variance in returns that cannot be diversified away. This result has great economic intuition for, if investors can eliminate all nonsystematic risk through diversification, there is no reason they should be rewarded, in terms of higher return, for bearing it. All of these implications of the CAPM are empirically testable. Indeed, in Chapter 15, we examine the results of these tests. Provided the tests hold, we have, with a simple model, gained great insight into the behavior of the capital markets.

We digress for a moment and point out one seeming fallacy in the potential use of the CAPM. Invariably, when a group of investors is first exposed to the CAPM, one or more investors will find a high-beta stock that last year produced a smaller return than low-beta stocks. The CAPM is an equilibrium relationship. High-beta stocks are expected to give a higher return than low-beta stocks because they are more risky. This does not mean that they will give a higher return over all intervals of time. In fact, if they always gave a higher return, they would be less risky, not more risky, than low-beta stocks. Rather, because they are more risky, they will sometimes produce lower returns. However, over long periods of time, they should on the average produce higher returns.

We have written the CAPM model in the form

$$\bar{R}_i = R_F + \beta_i\left(\bar{R}_M - R_F\right)$$

This is the form in which it is most often written and the form most amenable to empirical testing. However, there are alternative forms that give added insight into its meaning. Recall that

$$\beta_i = \frac{\sigma_{iM}}{\sigma_M^2}$$

We could then write the security market line as

$$\bar{R}_i = R_F + \left(\frac{\bar{R}_M - R_F}{\sigma_M}\right)\frac{\sigma_{iM}}{\sigma_M} \tag{13.3}$$

This, in fact, is the equation of a straight line located in expected return σ_{iM}/σ_M space. Recall that earlier in our discussion of the capital market, line $(\bar{R}_m - R_F)/\sigma_M$ was described as the market price of risk. Because σ_{iM}/σ_M is a definition of the risk of any security, or portfolio, we would see that the security market line, like the capital market line, states that

[6]This result is somewhat circular for, in this proof, we assumed that beta was the relevant risk measure. In the more rigorous proof that follows, we make no such assumption, yet we end up with the same equation for the security market line.

the expected return on any security is the riskless rate of interest plus the market price of risk times the amount of risk in the security or portfolio.[7]

Many authors write the CAPM equation as

$$\bar{R}_i = R_F + \left(\frac{\bar{R}_M - R_F}{\sigma_M^2} \right) \sigma_{iM}$$

They define $(\bar{R}_M - R_F)/\sigma_M^2$ as the market price of risk and σ_{iM} as the measure of the risk of security *i*. We have chosen the form we used because σ_{iM}/σ_M is the measure of how the risk on a security affects the risk of the market portfolio. It seems to us that this is the appropriate way to discuss the risk of a security.

We have now completed our intuitive proof of the CAPM. We are about to present a more complex mathematical proof. There are two reasons for presenting this mathematical proof. The first is that it is more rigorous. The second, and more important, reason is that one needs a richer framework to incorporate modifications of the assumptions of the standard CAPM. The method of proof just presented is too restrictive to allow forms of general equilibrium equations that make more realistic assumptions about the world to be derived. The framework presented subsequently can be used to derive equilibrium models under alternative assumptions and, indeed, will be used to do so in the next chapter. The reader who finds both these reasons unappealing can skip the next section and the derivations in the next chapter with no loss of continuity.

---

[7]In the following we offer theoretical justification that σ_{iM}/σ_M is the relevant measure of the risk of any security in equilibrium. Recall that the standard deviation of the market portfolio is given by

$$\sigma_M = \left[\sum_{i=1}^{N} X_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \sigma_{ij} \right]^{1/2}$$

where all X_is are market proportions. Because all investors hold the market portfolio, the relevant definition of the risk of a security is the change in the risk of the market portfolio, as the holdings of that security are varied. This can be found as follows:

$$\frac{d\sigma_M}{dX_i} = \frac{d\left[\sum_{i=1}^{N} X_i^2 \sigma_i^2 + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} X_i X_j \sigma_{ij} \right]^{1/2}}{dX_i}$$

$$= \frac{\left(\frac{1}{2}\right)\left[2X_i \sigma_i^2 + (2) \sum_{\substack{j=1 \\ j \neq 1}}^{N} X_j \sigma_{ij} \right]}{\left[\sum_{i=1}^{N} X_i^2 \sigma_i^2 + \sum_{j=1}^{N} \sum_{\substack{j=1 \\ j \neq 1}}^{N} X_i X_j \sigma_{ij} \right]^{1/2}} = \frac{X_i^2 \sigma_i^2 + \sum_{\substack{j=1 \\ j \neq 1}}^{N} X_j \sigma_{ij}}{\sigma_M} = \frac{\sigma_{iM}}{\sigma_M}$$

Therefore the relevant risk of security is equal to σ_{iM}/σ_M.

Deriving the CAPM—A More Rigorous Approach

To derive the CAPM more rigorously, we return to the analysis presented in Chapter 6. Recall that in the first section of Chapter 6, we solved for the optimal portfolio when short sales were allowed and the investor could lend and borrow unlimited amounts of money at the riskless rate of interest. The solution involved finding the composition of the portfolio that maximized the slope of a straight line passing through the riskless rate of interest on the vertical axes and the portfolio itself. As shown in Chapter 6, this involved maximizing the function

$$\theta = \frac{\overline{R}_P - R_F}{\sigma_P}$$

When the derivative of θ was taken with respect to all securities in the portfolio and each equation was set equal to zero, a set of simultaneous equations of the following form was derived:

$$\lambda\left(X_1\sigma_{1k} + X_2\sigma_{2k} + \cdots + X_k\sigma_k^2 + \cdots + X_N\sigma_{Nk}\right) = \overline{R}_k - R_F \tag{13.4}$$

This equation held for each security, and there is one such equation for each security in the market. If there are homogeneous expectations, then all investors must select the same optimum portfolio. If all investors select the same portfolio, then, in equilibrium, that portfolio must be a portfolio in which all securities are held in the same percentage that they represent of the market. In other words, in equilibrium, the proportion invested in security 1 must be that fraction of the total market value of all securities that security 1 represents. To get from Equation (13.4) to the CAPM involves simply recognizing that the left-hand side of Equation (13.4) is $\lambda \operatorname{cov}(R_k R_M)$. To see this, first note that

$$R_M = \sum_{i=1}^{N} R_i X_i'$$

where the prime indicates market proportions. Thus

$$\operatorname{cov}\left(R_k R_M\right) = E\left[\left(R_k - \overline{R}_k\right)\left(\sum_{i=1}^{N} R_i X_i' - \sum_{i=1}^{N} \overline{R}_i X_i'\right)\right] \tag{13.5}$$

Rearranging the second term,

$$\operatorname{cov}\left(R_k R_M\right) = E\left[\left(R_k - \overline{R}_k\right)\left(\sum_{i=1}^{N} X_i'\left(R_i - \overline{R}_i\right)\right)\right]$$

Multiplying out the terms,

$$\begin{aligned}\operatorname{cov}\left(R_k R_M\right) = E\Big[&X_1'\left(R_k - \overline{R}_k\right)\left(R_1 - \overline{R}_1\right) \\ &+ X_2'\left(R_k - \overline{R}_k\right)\left(R_2 - \overline{R}_2\right) + \cdots \\ &+ X_k'\left(R_k - \overline{R}_k\right)\left(R_k - \overline{R}_k\right) + \cdots + X_N'\left(R_k - \overline{R}_k\right)\left(R_N - \overline{R}_N\right)\Big]\end{aligned}$$

Because the expected value of the sum of random variables is the sum of the expected values, factoring out the *X*s yields

$$\begin{aligned}\operatorname{cov}(R_k R_M) = {} & X_1' E\left(R_k - \overline{R}_k\right)\left(R_1 - \overline{R}_1\right) + X_2' E\left(R_k - \overline{R}_k\right)\left(R_2 - \overline{R}_2\right) + \cdots \\ & + X_k' E\left(R_k - \overline{R}_k\right)^2 + \cdots + X_N' E\left(R_k - \overline{R}_k\right)\left(R_N - \overline{R}_N\right)\end{aligned}$$

Earlier we argued that the *X*s in Equation (13.4) were market proportions. Comparing Equation (13.5) with the left-hand side of Equation (13.4) shows that they are, indeed, equal. Thus Equation (13.4) can be written as

$$\lambda \operatorname{cov}(R_k R_M) = \overline{R}_k - R_F \tag{13.6}$$

Because this must hold for all securities (all possible values of *k*), it must hold for all portfolios of securities. One possible portfolio is the market portfolio. Writing Equation (13.6) for the market portfolio involves recognizing that $\operatorname{cov}(R_M R_M) = \sigma_M^2$:

$$\lambda \sigma_M^2 = \overline{R}_M - R_F$$

or

$$\lambda = \frac{\overline{R}_M - R_F}{\sigma_M^2}$$

Substituting this value for λ in Equation (13.6) and rearranging yields

$$\overline{R}_k = R_F + \frac{\overline{R}_M - R_F}{\sigma_M^2} \operatorname{cov}(R_k R_M) = R_F + \beta_k \left(\overline{R}_M - R_F\right)$$

This completes the more rigorous derivation of the security market line.

The advantages of this proof over that presented earlier are that we have not had to assume that beta is the relevant measure of risk, and we have established a framework that, as we see in the next chapter, can be used to derive general equilibrium solutions when some of the present assumptions are relaxed.

PRICES AND THE CAPM

Up to now, we have discussed equilibrium in terms of rate of return. In the introduction to this chapter, we mentioned that the CAPM could be used to describe equilibrium in terms of either return or prices. The latter is of importance in certain situations, for example, the pricing of new assets. It is very easy to move from the equilibrium relationship in terms of rates of return to one expressed in terms of prices. All that is involved is a little algebra.

Let us define

P_i as the present price of asset *i*.

P_M as the present price of the market portfolio (all assets).

Y_i as the dollar value of the asset one period hence. It is market value plus any dividends.

Y_M as the dollar value of the market portfolio one period hence, including dividends.

$\operatorname{cov}(Y_i Y_M)$ as the covariance between Y_i and Y_M.

$\operatorname{var}(Y_M)$ as the variance in Y_M.

r_F as $(1 + R_F)$.

The return on asset i is

$$R_i = \frac{\text{Ending value} - \text{Beginning value}}{\text{Beginning value}}$$

In symbols,

$$R_i = \frac{Y_i - P_i}{P_i} = \frac{Y_i}{P_i} - 1$$

Similarly,

$$R_M = \frac{Y_M - P_M}{P_M} = \frac{Y_M}{P_M} - 1$$

Substituting these expressions into Equation (13.3) yields

$$\frac{\overline{Y}_i}{P_i} - 1 = R_F + \left(\frac{\overline{Y}_M}{P_M} - 1 - R_F\right)\frac{\text{cov}(R_i R_M)}{\sigma_M^2} \tag{13.7}$$

Now we can rewrite $\text{cov}(R_i R_M)$ as

$$\begin{aligned}\text{cov}(R_i R_M) &= E\left[\left(\frac{Y_i - P_i}{P_i} - \frac{\overline{Y}_i - P_i}{P_i}\right)\left(\frac{Y_M - P_M}{P_M} - \frac{\overline{Y}_M - P_M}{P_M}\right)\right] \\ &= E\left[\left(\frac{Y_i}{P_i} - \frac{\overline{Y}_i}{P_i}\right)\left(\frac{Y_M}{P_M} - \frac{\overline{Y}_M}{P_M}\right)\right] = \frac{1}{P_i P_M}\text{cov}(Y_i Y_M)\end{aligned}$$

Similarly,

$$\sigma_M^2 = \frac{1}{P_M^2}\text{var}(Y_M)$$

Substituting these into Equation (13.7), adding 1 to both sides of the equation, and recalling that $r_F = 1 + R_F$,

$$\frac{\overline{Y}_i}{P_i} = r_F + \left(\frac{\overline{Y}_M}{P_M} - r_F\right)\frac{\frac{1}{P_i}\frac{1}{P_M}\text{cov}(Y_i Y_M)}{\frac{1}{P_M^2}\text{var}(Y_M)}$$

Multiplying both sides of the equation by P_i and simplifying the last term on the right-hand side,

$$\overline{Y}_i = r_F P_i + \left(\overline{Y}_M - r_F P_M\right)\frac{\text{cov}(Y_i Y_M)}{\text{var}(Y_M)}$$

Solving this expression for P_i,

$$P_i = \frac{1}{r_F}\left[\overline{Y}_i - \left(\overline{Y}_M - r_F P_M\right)\frac{\text{cov}(Y_i Y_M)}{\text{var}(Y_M)}\right]$$

Valuation formulas of this type have often been suggested in the security analysis literature. The equation involves taking the expected dollar return next year, ($\bar{Y}_i$), subtracting off some payment as compensation for risk taking, and then taking the present value of the net result. The term in square brackets can be thought of as the certainty equivalent of the horizon cash payment, and to find the present value of the certainty equivalent, we simply discount it at the riskless rate of interest. Although this general idea is not new, the explicit definition of how to find the certainty equivalent is one of the fundamental contributions of the CAPM. It can be shown that

$$\frac{\bar{Y}_M - r_F P_M}{\left[\text{var}(Y_M)\right]^{1/2}}$$

is equal to a measure of the market price of risk and that

$$\frac{\text{cov}(Y_i Y_M)}{\left[\text{var}(Y_M)\right]^{1/2}}$$

is the relevant measure of risk for any asset.

CONCLUSION

In this chapter we have discussed the Sharpe–Lintner–Mossin form of a general equilibrium relationship in the capital markets. This model, usually referred to as the capital asset pricing model or standard CAPM, is a fundamental contribution to understanding the manner in which capital markets function. It is worthwhile highlighting some of the implications of this model.

First, we have shown that, under the assumptions of the CAPM, the only portfolio of risky assets that any investor will own is the market portfolio. Recall that the market portfolio is a portfolio in which the fraction invested in any asset is equal to the market value of that asset divided by the market value of all risky assets. Each investor will adjust the risk of the market portfolio to her preferred risk-return combination by combining the market portfolio with lending or borrowing at the riskless rate. This leads directly to the two mutual fund theorem. The two mutual fund theorem states that all investors can construct an optimum portfolio by combining a market fund with the riskless asset. Thus all investors will hold a portfolio along the line connecting R_F with $\bar{R}_M$ in expected return, standard deviation of return space. See Figure 13.5.

This line, usually called the capital market line, which describes all efficient portfolios, is a pictorial representation of the equation

$$\bar{R}_e = R_F + \frac{\bar{R}_M - R_F}{\sigma_M}\sigma_e$$

Thus we can say that the return on an efficient portfolio is given by the market price of time plus the market price of risk times the amount of risk on an efficient portfolio. Note that risk is defined as the standard deviation of return on any efficient portfolio.

From the equilibrium relationship for efficient portfolios we were able to derive the equilibrium relationship for any security or portfolio (efficient or inefficient). This relationship, presented in Figure 13.6, is given by

$$\bar{R}_i = R_F + \left(\frac{\bar{R}_M - R_F}{\sigma_M}\right)\frac{\sigma_{iM}}{\sigma_M}$$

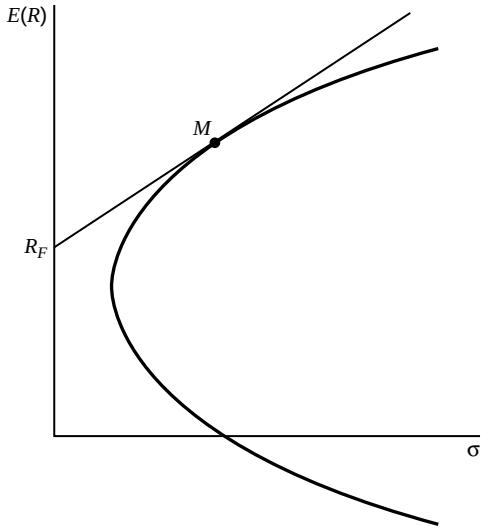


Figure 13.5 The efficient frontier.

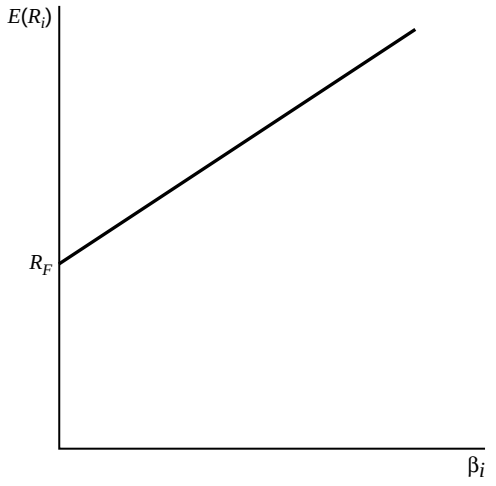


Figure 13.6 The security market line.

or

$$\overline{R}_i = R_F + \beta_i\left(\overline{R}_M - R_F\right)$$

This relationship is usually called the security market line. Notice that it might have been called the security-portfolio market line, for it describes the equilibrium return on all portfolios as well as all securities.

Examination of the first form of the security market line shows that it is analogous in many ways to the capital market line. As we have shown, the impact of a security on the risk of the market portfolio is given by σ_{iM}/σ_M. Thus we can state that the equilibrium return on any security is equal to the price of time plus the market price of risk times the relevant definition of risk for the security.

The security market line clearly shows that return is an increasing function, in fact, a linearly increasing function, of risk. Furthermore, it is only market risk that affects return. The investor receives no added return for bearing diversifiable risk.

The capital asset pricing model has been derived under a set of very restrictive assumptions. The test of a model is how well it describes reality. The key test is: Does it describe the behavior of returns in the capital markets? These tests will be taken up in Chapter 15. Before we turn to these tests, however, it is logical to examine forms of the general equilibrium relationship that exist under less restrictive assumptions. Even if the standard CAPM model explains the behavior of security returns, it obviously does not explain the behavior of individual investors. Individual investors hold nonmarket and, in fact, quite often, very small portfolios. Furthermore, by developing alternative forms of the general equilibrium relationship, we can test whether observed returns are more consistent with one of these than they are with the standard CAPM.

APPENDIX

Appropriateness of the Single-Period Asset Pricing Model

Up to now, we have assumed that all investors make investment decisions based on a single-period horizon. In fact, the portfolio an investor selects, at any point in time, is really one step in a series of portfolios that he intends to hold over time to maximize his utility of lifetime consumption. Two questions immediately become apparent:

1. What are the conditions under which the simple CAPM adequately describes market equilibrium?
2. Is there a fully general multiperiod equilibrium model?

Fama (1970) and Elton and Gruber (1974, 1975) have explored the conditions under which the multiperiod investment consumption decision can be reduced to the problem of maximizing a one-period utility function. These conditions are as follows:

1. The consumer's tastes for particular consumption goods and services are independent of future events (any future sets of conditions).
2. The consumer acts as if consumption opportunities in terms of goods and their prices are known at the beginning of the decision period (are not state dependent).
3. The consumer acts as if the distribution of one-period returns on all assets are known at the beginning of the decision period (are not state dependent).

Hansen and Jagannathan (1991) have developed a very simple and elegant approach to developing a multiperiod investment consumption equilibrium model based on these assumptions, an approach that builds on earlier work by Breeden (1999) and Rubinstein (1974). The investor's problem is to allocate wealth to maximize the utility of consuming both now and in the future. In other words, the investor is faced with an intertemporal choice problem. Maximize the expected value of the present value of future consumption,

$$\text{Max } E_t\left[\sum_{j=0}^{\infty} \delta^j U(c_{t+j})\right]$$

where c_{t+j} represents future consumption and δ^j is a subjective discount factor applied to future consumption in period j. For a given budget constraint, the first-order conditions for this problem imply that

$$U'(c_t) = \delta^j E_t\left[(1 + R_{i,t+j}) U'(c_{t+j})\right]$$

for all assets i and periods j into the future. Dividing through by the marginal utility of consumption today, we have the important result that

$$1 = E_t\left[(1 + R_{i,t+j}) m_{t,j}\right]$$

where $m_{t,j} = \delta^j[U'(c_{t+j}) / U'(c_t)]$ is the intertemporal marginal rate of substitution. It also has the interesting interpretation of being a *stochastic discount factor* (sometimes also referred to as a *pricing kernel*) because it takes an asset with uncertain per dollar future payoff back to the present to be valued at \$1. If there is a riskless asset in this economy with return $R_{F,t+j}$, then $1 = E_t[(1 + R_{F,t+j})m_{t,j}] = (1 + R_{F,t+j})E_t[m_{t,j}]$ or $E_t[m_{t,j}] = 1 / (1 + R_{F,t+j})$, so that the expected value of the stochastic discount factor is equal to the discount factor used when the future payment is in fact without any risk. For the subsequent discussion, we will drop the time subscripts.

Cochrane (2001) argues that this is a straightforward way to value all financial claims. The difficulty is, however, that the stochastic discount factor m is not observable. There are three general approaches to this problem. The first is to specify m directly through assumptions made about utility and using measures of consumption. The chief difficulty associated with this approach is obtaining accurate and timely measures of aggregate consumption c. An alternative approach is to use a vector of factors, some combination of which can proxy for consumption growth. A third idea, originally from Hansen and Jagannathan (1991), is that because the stochastic discount factor prices all financial claims, we might be just as well off inferring the stochastic discount factor from the observed set of asset returns. This important insight allows us to interpret the stochastic discount factor in terms of the mean–variance efficient portfolio, an interpretation that yields the standard CAPM as a direct implication.

The relationship between the stochastic discount factor and the mean–variance efficient portfolio is quite direct. Starting from the basic formula that holds for all securities i,

$$\begin{aligned} 1 &= E\left[(1 + R_i) m\right] \\ &= E\left[(1 + \bar{R}_i + (R_i - \bar{R}_i) m)\right] \\ &= (1 + \bar{R}_i) E[m] + E\left[(R_i - \bar{R}_i) m\right] \end{aligned}$$

If there is a risk-free rate $E[m] = 1 / (1 + R_F)$, this implies the asset pricing relation $\bar{R}_i - R_F = -(1 + R_F) E[(R_i - \bar{R}_i)m]$, which says that the risk premium is negatively proportional to the covariance between the asset return and the stochastic discount factor. In difficult economic times, consumption is depressed, and the intertemporal rate of substitution m is high. A negative covariance between asset returns and m is therefore a source of risk to investors.

As we mention earlier, we cannot directly observe m. This is, of course, a serious challenge to empirical application of this theory. However, the same m prices all assets in the economy, and so we should be able to infer it from asset prices and returns. If asset returns and the (unobserved) stochastic discount factor are multivariate normal (this assumption can be generalized to include fat-tailed alternatives to normality such as multivariate Student t, stable, and other distribution functions), the conditional expectation of the stochastic discount factor is linear in asset returns. Another way of saying this is that if there

are N assets in the economy, we can infer (the unobserved) m by a hypothetical regression on the set of returns in the economy, so that

$$m = \overline{m} + \sum_{j=1}^{N} (R_j - \overline{R}_j)\gamma_j$$

where the γ_j are the hypothetical regression coefficients. The asset pricing relationship can then be written as follows:

$$\begin{aligned}
\overline{R}_i - R_F &= -(1+R_F)\,E\left[(R_i - \overline{R}_i)\,m\right] \\
&= -(1+R_F)\,\overline{m}\,E\left[(R_i - \overline{R}_i)\right] - (1+R_F)\,E\left[(R_i - \overline{R}_i)\sum_{j=1}^{N}(R_j - \overline{R}_j)\gamma_j\right] \\
&= \sum_{j=1}^{N} E(R_i - \overline{R}_i)\,(R_j - \overline{R}_j)\,Z_j
\end{aligned}$$

where the coefficients $Z_j = -(1+R_F)\gamma_j$. This equation must hold for all assets, so we have the system of equations

$$\begin{aligned}
\overline{R}_i - R_F &= Z_1\sigma_1^2 + Z_2\sigma_{12} + Z_3\sigma_{13} + \ldots + Z_N\sigma_{1N} \\
\overline{R}_2 - R_F &= Z_1\sigma_{21} + Z_2\sigma_2^2 + Z_3\sigma_{23} + \ldots + Z_N\sigma_{2N} \\
\overline{R}_3 - R_F &= Z_1\sigma_{31} + Z_2\sigma_{32} + Z_3\sigma_3^2 + \ldots + Z_N\sigma_{3N} \\
&\vdots \\
\overline{R}_N - R_F &= Z_1\sigma_{N1} + Z_2\sigma_{N2} + Z_3\sigma_{N3} + \ldots + Z_N\sigma_N^2
\end{aligned}$$

which the reader will recognize as Equation (6.1), used to identify the mean–variance efficient portfolio with riskless lending and borrowing. From this fact, we conclude immediately that the hypothetical regression coefficients γ_j are proportional to mean–variance efficient portfolio weights, and hence that the best estimate of the stochastic discount factor m^* can be written as a linear function of the return on a mean–variance efficient portfolio, $m^* = a + bR_{MV}$. If we further identify this portfolio as the market portfolio, then the previous asset pricing relation immediately implies the standard CAPM.

To see this, note that, using this proxy for the discount factor,

$$\begin{aligned}
\overline{R}_i - R_F &= -(1+R_F)\,E\left[(R_i - \overline{R}_i)\,m^*\right] \\
&= -(1+R_F)\,a\,E\left[(R_i - \overline{R}_i)\right] - b(1+R_F)\,E\left[(R_i - \overline{R}_i)R_{MV}\right] \\
&= -b(1+R_F)\sigma_{MV}^2\beta_i
\end{aligned}$$

which must hold for all assets, including the mean–variance efficient portfolio with beta equal to one, so that $b = -(\overline{R}_{MV} - R_F) / [(1 + R_F)\sigma_{MV}^2]$ and $\overline{R}_i - R_F = \beta_i(\overline{R}_{MV} - R_F)$. There is an important intuition here that establishes that the particular mean–variance efficient portfolio is the market portfolio. From the interpretation of the portfolio as resulting from regressing the (unknown) stochastic discount factor m on the set of observed security returns, the variability explained by the observed return portfolio, $\sigma_{m^*}^2 = b^2\sigma_{MV}^2 = [(\overline{R}_{MV} - R_F)^2/(1 + R_F)^2\sigma_{MV}^2]$, is maximized. This implies that the Sharpe ratio given as $(R_{MV} - R_F)/\sigma_{MV} = (1 + R_F)\sigma_{m^*} = (\sigma_{m^*} / \overline{m}^*)$ is also maximized, which further establishes that R_{MV} is the return on the market portfolio.

This last equation is very important because it establishes a nexus between the financial markets on one hand and consumer preferences on the other. The fact that we can represent the maximized Sharpe ratio as the ratio of the standard deviation of the stochastic discount factor to its mean presents many financial economists with a serious problem. The average risk premium of the market measured in units of risk is far too high to be explained by any consumption-based representation of the stochastic discount factor. This is referred to as the *equity premium puzzle*. As Cochrane (2001) notes, the Sharpe ratio measured in real (not nominal) terms has been about 0.5 on the basis of the past 50 years of data for the United States. Assuming the time separable power utility function (Chapter 11), the ratio $\sigma_{m^*} / \bar{m}^*$ is approximately the risk-aversion coefficient times the standard deviation of the log of consumption. Because the rate of change in consumption is considerably less than the variance of market returns, this implies that investors are very risk averse, with a coefficient of risk aversion at least 50. A degree of risk aversion this large is difficult to motivate.

The relationship between the Sharpe ratio and the moments of the stochastic discount factor gives rise to another fascinating insight by Hansen and Jagannathan (1991). Suppose we represent the stochastic discount factor $m^* = a + bR_{MV}$. We have shown that this choice of m^* prices all assets, $E[(1+R)m^*]=1$. Consider another discount factor $m = a + bR_{MV} + \varepsilon$, where ε is uncorrelated with returns R on every single asset in the economy and has zero expectation. Then this discount factor will also price all assets $E[(1 + R)m] = 1$. This means that there are many possible discount factors, with $\bar{m} = \bar{m}^*$ and $\sigma_m^2 = \sigma_{m^*}^2 + \sigma_\varepsilon^2 > \sigma_{m^*}^2$. However, the choice of $m = m^*$ minimizes the volatility of the stochastic discount factor and is most preferred by investors. Hence we have the interesting bounds: $\sigma_m / \bar{m} \geq \sigma_{m^*} / \bar{m}^* = (\bar{R}_{MV} - R_F) / \sigma_{MV} \geq (\bar{R} - R_F) / \sigma_{MV}$, which are referred to as the *Hansen-Jagannathan bounds*. The inequality on the left-hand side refers to the fact that consumption risk will increase if there is a source of risk ε that cannot be hedged by the set of assets represented by returns R. Imperfections in the capital markets make the world a riskier place than it needs to be. The inequality on the right-hand side reflects the reality that limits to diversification through short sale restrictions or other factors limit the opportunities available to investors. Another useful interpretation of these contrasting inequalities is that the problem of choosing a returns-based discount function by minimizing the volatility of the discount factor m is equivalent to determining a portfolio that maximizes the portfolio Sharpe ratio.

Unfortunately, this interpretation of the stochastic discount factor leads to a distinctly unattractive implication. Because $\bar{m}^* = 1 / (1 + R_f) = a + b\bar{R}_{MV}$, $a = 1 / (1 + R_f) + (\bar{R}_{MV} - R_F) / [(1 + R_F)\sigma_{MV}^2]\bar{R}_{MV}$, and therefore $m^* = 1 / (1 + R_f) - [(\bar{R}_{MV} - R_F) / (1 + R_F)\sigma_{MV}^2]\ (R_{MV} - \bar{R}_{MV})$, so that the implied discount rate is negative whenever the market return exceeds its mean by an amount equal to $\sigma_{MV}^2 / (\bar{R}_{MV} - R_F)$, the standard deviation of the market return divided by the Sharpe ratio. Using the preceding example, if the Sharpe ratio is about 0.5, the discount factor will be negative whenever the market return is 2 standard deviations above its mean. It is intuitive that the discount factor should fall as market return increases; after all, we are assuming that consumers have diminishing marginal utility. It is not intuitive that the discount factor should be negative, and in fact it is easy to show that there are arbitrage opportunities that arise when the representative investor is willing to throw his money away in this way.

As a practical matter, the empirical representation of the discount factor in terms of the return on the market portfolio is rarely negative, and so a constraint on the discount factor to ensure that it is nonnegative should not lead to any major implications for asset prices. Hansen and Jagannathan suggest such a restriction but find it makes little difference in the

case of equity markets. This assertion, however, depends strongly on the assumption that market returns are normally distributed. If returns are positively skewed, then the positive discount factor restriction may have a greater impact.

It is worth pursuing a little the implications of a positive discount factor restriction. We can represent this positive factor as $m^+ = a + b(R_{MV} - c)^-$, where $(R_{MV} - c)^-$ is the payoff of a put on the market index with exercise price $1 + c$. As before, we have

$$\begin{aligned}\bar{R}_i - R_F &= -(1+R_F)\,E\left[(R_i - \bar{R}_i)\,m^+\right] \\ &= -(1+R_F)\,a\,E\left[(R_i - \bar{R}_i)\right] - b(1+R_F)\,E\left[(R_i - \bar{R}_i)(R_{MV} - c)^-\right. \\ &= -b(1+R_F)\,LPM_{MV,c}\,\beta^-_{MV,c,i}\end{aligned}$$

where $LPM_{MV,c} = E[(R_{MV} - c)^2 \mid R_{MV} < c]$ is referred to as the lower partial moment of the return on the mean–variance efficient portfolio, a measure of downside risk, and where $\beta^-_{MV,c,i} = E\left[(R_i - \bar{R}_i)(R_{MV} - c)^-\right] / LPM_{MV}$ is referred to as the lower partial moment beta, the contribution of security *i* to the downside risk of the market. As before, we have immediately a linear asset pricing model similar to the standard CAPM, except that the lower partial moment beta, $\beta^-_{MV,c,i}$, replaces the more familiar beta, β_{MV}. This generalized asset pricing model was first derived by Bawa and Lindenberg (1977) who showed that it corresponded with an equilibrium model where agents have utility functions for wealth displaying diminishing absolute risk aversion. They further show that if returns are multivariate normal or Student *t*, the $\beta^-_{MV,c,i}$ risk measure collapses to β_{MV}, and the standard CAPM result follows.

QUESTIONS AND PROBLEMS

1. Assume that the following assets are correctly priced according to the security market line. Derive the security market line. What is the expected return on an asset with a beta of 2?

$$\begin{aligned}\bar{R}_1 &= 6\% \quad \beta_1 = 0.5 \\ \bar{R}_2 &= 12\% \quad \beta_2 = 1.5\end{aligned}$$

2. Assume the security market line given below. Assume that analysts have estimated the beta on two stocks as follows: $\beta_x = 0.5$ and $\beta_y = 2$. What must the expected return on the two securities be in order for them to be a good purchase?

$$\bar{R}_i = 0.04 + 0.08\beta_i$$

3. Assume that over some period, a CAPM was estimated. The results are shown below. Assume that over the same period, two mutual funds had the following results:

Fund A Actual return = 10% Beta = 0.8
Fund B Actual return = 15% Beta = 1.2

What can be said about the fund performance?

$$\bar{R}_i = 0.06 + 0.19\beta_i$$

4. Consider the CAPM line shown below. What is the excess return of the market over the risk-free rate? What is the risk-free rate?

$$\bar{R}_i = 0.04 + 0.10\beta_i$$

5. Write the CAPM shown in Problem 4 in price form.

6. Show that the standard CAPM should hold even if short sales are not allowed.

7. Assume that an asset exists with $\bar{R}_3 = 15\%$ and $\beta_3 = 1.2$. Further assume the security market line discussed in Problem 1. Design the arbitrage opportunity.

8. If the following assets are correctly priced on the security market line, what is the return of the market portfolio? What is the risk-free rate?

$$\bar{R}_1 = 9.40\% \quad \beta_1 = 0.80$$
$$\bar{R}_2 = 13.40\% \quad \beta_2 = 1.30$$

9. Given the security market line

$$\bar{R}_i = 0.07 + 0.09\beta_i$$

What must be the returns for two stocks, assuming their βs are 1.2 and 0.9?

BIBLIOGRAPHY

1. Aivazian, Varouj. "The Demand for Assets under Conditions of Risk: Comment," *Journal of Finance*, **XXXII,** No. 3 (June 1976), pp. 927–929.
2. Bawa, Vijay, and Lindenburg, Eric. "Capital Market Equilibrium in a Mean-Lower Partial Moment Framework," *Journal of Financial Economics*, **5** (1977), pp. 189–200.
3. Benninga, Simon, and Protopapadakis, Aris. "The Stock Market Premium, Production, and Relative Risk Aversion," *The American Economic Review*, **81,** No. 3 (June 1991), pp. 591–599.
4. Bernstein, Peter L. "What Rate of Return Can You 'Reasonably' Expect?" *Journal of Finance*, **XXVIII,** No. 2 (May 1973), pp. 273–282.
5. Breeden, Douglas. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, **7** (1999), pp. 265–296.
6. Chen, Nai-Fu, Grundy, Bruce, and Stambaugh, Robert F. "Changing Risk, Changing Risk Premiums, and Dividend Yield Effects," *The Journal of Business*, **63,** No. 1 (Jan. 1990), pp. 551–570.
7. Cochrane, John. *Asset Pricing* (Princeton, NJ: Princeton University Press, 2001).
8. Elton, Edwin J., and Gruber, Martin J. "The Multi-period Consumption Investment Decision and Single-Period Analysis," *Oxford Economic Paper*, **26** (Sept. 1974), pp. 180–195.
9. Elton, Edwin J., and Gruber, Martin J. *Finance as a Dynamic Process* (Englewood Cliffs, NJ: Prentice Hall, 1975).
10. Fama, Eugene. "Risk, Return and Equilibrium: Some Clarifying Comments," *Journal of Finance*, **XXIII,** No. 1 (March 1968), pp. 29–40.
11. Fama, Eugene. "Multi-period Consumption Investment Decision," *American Economic Review*, **60** (March 1970), pp. 163–174.
12. ——. "Risk, Return and Equilibrium," *Journal of Political Economy*, **79,** No. 1 (Jan./Feb. 1971), pp. 30–55.
13. ——. "Risk, Return and Portfolio Analysis: Reply to [20]," *Journal of Political Economy*, **81,** No. 3 (May/June 1973), pp. 753–755.
14. Fama, Eugene F. "Determining the Number of Priced State Variables in the ICAPM," *Journal of Financial and Quantitative Analysis*, **33,** No. 2 (June 1998), pp. 217–231.
15. Ferson, Wayne E., Harvey, C., and Campbell, R. "The Variation of Economic Risk Premiums," *The Journal of Political Economy*, **99,** No. 2 (April 1991), pp. 385–416.
16. Ferson, Wayne E., Kandel, Shmuel, and Stambaugh, Robert F. "Tests of Asset Pricing with Time-Varying Expected Risk Premiums and Market Betas," *The Journal of Finance*, **42,** No. 2 (June 1987), pp. 201–220.
17. Green, Richard C. "Benchmark Portfolio Inefficiency and Deviations from the Security Market Line," *The Journal of Finance*, **41,** No. 2 (June 1986), pp. 295–312.

18. Hansen, Lars Peter, and Jagannathan, Ravi. "Implications of Security Market Data for Models of Dynamic Economics," *Journal of Political Economy*, **99** (1991), pp. 225–262.
19. Hietala, Pekka T. "Asset Pricing in Partially Segmented Markets: Evidence from the Finnish Market," *The Journal of Finance*, **44,** No. 3 (July 1989), pp. 697–718.
20. Kroll, Yoram, and Levy, Haim. "Further Tests of the Separation Theorem and the Capital Asset Pricing Model," *The American Economic Review*, **82,** No. 3 (June 1992), pp. 664–670.
21. Kroll, Yoram, Levy, Haim, and Rapoport, Amnon. "Experimental Tests of the Separation Theorem and the Capita," *The American Economic Review*, **78,** No. 3 (June 1988), pp. 500–519.
22. Levy, Haim. "The Demand for Assets under Conditions of Risk," *Journal of Finance*, **XXVIII,** No. 1 (March 1973), pp. 79–96.
23. ——. "The Demand for Assets under Conditions of Risk: Reply to [1]," *Journal of Finance*, **XXXII,** No. 3 (June 1976), pp. 930–932.
24. Lintner, John. "Security Prices, Risk, and Maximal Gains from Diversification," *Journal of Finance* (Dec. 1965), pp. 587–615.
25. ——. "The Aggregation of Investor's Diverse Judgments and Preferences in Purely Competitive Security Markets," *Journal of Financial and Quantitative Analysis*, **IV,** No. 4 (Dec. 1969), pp. 347–400.
26. ——. "The Market Price of Risk, Size of Market and Investor's Risk Aversion," *Review of Economics and Statistics*, **LII,** No. 1 (Feb. 1970), pp. 87–99.
27. Markowitz, Harry M. "Nonnegative or Not Nonnegative: A Question about CAPMs," *The Journal of Finance*, **38,** No. 2 (May 1983), pp. 283–296.
28. Modigliani, Franco, and Pogue, Jerry. "An Introduction to Risk and Return," *Financial Analysts Journal*, **30,** No. 2 (Mar./Apr. 1974), pp. 68–80.
29. ——. "An Introduction to Risk and Return: Part II," *Financial Analysts Journal*, **30,** No. 3 (May/June 1974), pp. 69–86.
30. Mossin, Jan. "Equilibrium in a Capital Asset Market," *Econometrica*, **34** (Oct. 1996), pp. 768–783.
31. Ng, Lilian. "Tests of the CAPM with Time-Varying Covariances: A Multivariate GARCH Approach," *The Journal of Finance*, **46,** No. 4 (Sept. 1991), pp. 1507–1521.
32. Ross, Stephen. "A Simple Approach to the Valuation of Risky Streams," *Journal of Business*, **51,** No. 3 (July 1978), pp. 453–475.
33. Rubinstein, Mark. "An Aggregation Theorem for Securities Markets," *Journal of Financial Economy*, **1,** No. 3 (Sept. 1974), pp. 225–244.
34. Rubinstein, Mark E. "A Mean-Variance Synthesis of Corporate Financial Theory," *Journal of Finance*, **XXXVIII,** No. 1 (March 1973), pp. 167–181.
35. Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance* (Sept. 1964), pp. 425–442.
36. ——. "Bonds versus Stocks: Some Lessons from Capital Market Theory," *Financial Analysts Journal*, **29,** No. 6 (Nov./Dec. 1973), pp. 74–80.
37. ——. "Capital Asset Prices with and without Negative Holdings," *The Journal of Finance*, **46,** No. 2 (June 1991), pp. 489–509.
38. Stapleton, C. Richard. "Portfolio Analysis, Stock Valuation and Capital Budgeting Decision Rules for Risky Projects," *Journal of Finance*, **XXVI,** No. 1 (March 1971), pp. 95–117.
39. Tinic, Seha M., and West, Richard R. "Risk, Return, and Equilibrium: A Revisit," *The Journal of Political Economy*, **94,** No. 1 (Feb. 1986), pp. 126–147.
40. Tsiang, S. C. "Risk, Return and Portfolio Analysis: Comment on [4]," *Journal of Political Economy*, **81,** No. 3 (May/June 1973), pp. 748–752.
41. Turnbull, Stuart. "Market Value and Systematic Risk," *Journal of Finance*, **XXXII,** No. 4 (Sept. 1977), pp. 1125–1142.

14

Nonstandard Forms of Capital Asset Pricing Models

The Capital Asset Pricing Model (CAPM) model developed in the previous chapter would provide a complete description of the behavior of capital markets if each of the assumptions set forth held. The test of the CAPM model is how well it describes reality. But even before we examine these tests, it is useful to develop equilibrium models based on more realistic assumptions. Most of the assumptions underlying the CAPM violate conditions in the real world. This does not mean that we should disregard the CAPM model, for the differences from reality may be sufficiently unimportant that they do not materially affect the explanatory power of the model. On the other hand, the incorporation of alternative, more realistic assumptions into the model has several important benefits. Although the CAPM may describe equilibrium returns on the macro level, it certainly is not descriptive of micro (individual investor) behavior. For example, most individuals and many institutions hold portfolios of risky assets that do not resemble the market portfolio. We might get better insight into investor behavior by examining models developed under alternative and more realistic assumptions. Another reason for examining other equilibrium models is that it allows us to formulate and test alternative explanations of equilibrium returns. The CAPM may work well, but do other models work better and explain discrepancies from the CAPM? Finally, and perhaps most important, because the CAPM assumes several real-world influences away, it does not provide us with a mechanism for studying the impact of those influences on capital market equilibrium or on individual decision making. Only by recognizing the presence of these influences can their impact be investigated. For example, if we assume personal taxes do not exist, there is no way the equilibrium model can be used to study the effects of taxes. By constructing a model that includes taxes, we can study the impact of taxes on individual investor behavior and on equilibrium returns in the capital market.

The effects of modifying most of the assumptions of the CAPM model have been examined in the economics and finance literature. We review much of this work in this chapter. We place special emphasis on two assumptions: the ability to lend and borrow infinite sums of money at the riskless rate and the absence of personal taxes. The reason we do so is not only because there are important influences but also because they lead to the development of full-fledged general equilibrium models of a form that are amenable to testing.

In the remainder of this chapter we discuss general equilibrium models derived under more realistic assumptions about each of the following influences:

Short sales disallowed
Riskless lending and borrowing
Personal taxes
Nonmarketable assets
Heterogeneous expectations
Non-price-taking behavior
Multiperiod CAPM

SHORT SALES DISALLOWED

One of the assumptions made in deriving the CAPM is that the investor can engage in unlimited short sales. Furthermore, short sales were defined in the broadest sense of the term in that the investor was allowed to sell any security (whether owned or not) and to use the proceeds to buy any other security.[1] This was a convenient assumption and it simplified the mathematics of the derivation, but it was *not* a necessary assumption. Exactly the same result would have been obtained had short sales been disallowed. The economic intuition behind this is quite simple.[2] In the CAPM framework all investors hold the market portfolio in equilibrium. Because, in equilibrium, no investor sells any security short, prohibiting short selling cannot change the equilibrium.[3] Thus the same CAPM relationship would be derived irrespective of whether short sales are allowed or prohibited.

MODIFICATIONS OF RISKLESS LENDING AND BORROWING

A second assumption of the CAPM is that investors can lend and borrow unlimited sums of money at the riskless rate of interest. Such an assumption is clearly not descriptive of the real world. It seems much more realistic to assume that investors can lend unlimited sums of money at the riskless rate but cannot borrow at a riskless rate. The lending assumption is equivalent to investors being able to buy government securities equal in maturity to their single-period horizon. Such securities exist and they are, for all intents and purposes, riskless. Furthermore, the rate on such securities is virtually the same for all investors. On the other hand, it is not possible for investors to borrow unlimited amounts at a riskless rate. It is convenient to examine the case where investors can neither borrow nor lend at the riskless rate first, and then to extend the analysis to the case where they can lend but not borrow at the riskless rate.

---

[1]The allowance of short sales was reflected in the constraint on our basic problem in Chapter 6 that $\Sigma X_i = 1$, while simultaneously not constraining X_i to be positive.

[2]For a formal proof, see Lintner (1971).

[3]The more mathematically inclined reader can reach this same conclusion by using the Kuhn–Tucker conditions on the basic problem outlined in the previous chapter. The derivative of the Lagrangian with respect to each security will have a Kuhn–Tucker multiplier added to it but since each security is contained in the market portfolio, the value of each Kuhn–Tucker multiplier will be zero. Hence the solution will be unchanged.

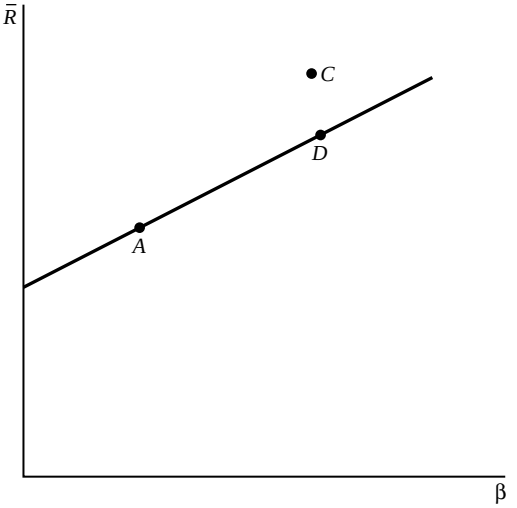


Figure 14.1 Portfolios in expected return beta space.

No Riskless Lending or Borrowing

This model is the second most widely used general equilibrium model. The simple capital asset pricing model developed in the last chapter is the most widely used. Because of the importance of this model, we derive it twice. The first derivation stresses economic rationale, the second is more rigorous.

Simple Proof In the last chapter we argued that systematic risk was the appropriate measure of risk and that two assets with the same systematic risk could not offer different rates of return. The essence of the argument was that the unsystematic risk of large diversified portfolios was essentially zero. Thus, even if an individual asset had a great deal of unsystematic risk, it would have little impact on portfolio risk, and therefore, unsystematic risk would not require a higher return. This was formalized in Figure 13.3, and an analogous diagram, Figure 14.1, will be used here.

Let us recall why all assets are plotted on a straight line. First, we showed that combinations of two risky portfolios lie on a straight line connecting them in expected return beta space. For example, positive combinations of portfolios *A* and *D* lie on the line segment *A–D*. Thus, if securities or portfolios happened to lie on a straight line in expected return beta space, all combinations of securities (e.g., portfolios) would lie on the same line.

Now consider securities *C* and *D* in Figure 14.1. They both have the same systematic risk, but *C* has a higher return. Clearly an investor would purchase *C* rather than *D* until the prices adjusted so that they offered the same return. In fact, an investor could purchase *C* and sell *D* short and have an asset with positive expected return and no systematic risk. Such an opportunity cannot exist in equilibrium. In short, all portfolios and securities must plot along a straight line.

One portfolio that lies along the straight line is the market portfolio. This can be seen in either of two ways. If it did not lie along the straight line, two assets would exist with the same systematic risk and different return, and in equilibrium, equivalent assets must offer the same return. In addition, note that all combinations of securities lie on the line and the market portfolio is a weighted average of the securities.

A straight line can be described by any two points. One convenient point is the market portfolio. A second convenient portfolio is where the straight line cuts the vertical axis (where beta equals zero).[4]

The equation of a straight line is

$$\text{Expected return} = a + b(\text{Beta})$$

This must hold for a portfolio with zero beta. Letting $\bar{R}_Z$ be the expected return on this portfolio, we have

$$\bar{R}_Z = a + b(0) \qquad \text{or} \qquad a = \bar{R}_Z$$

The equation must also hold for the market portfolio. If $\bar{R}_M$ is the expected return on the market and, recalling that the beta for the market portfolio is 1, we have

$$\bar{R}_M = \bar{R}_Z + b(1) \qquad \text{or} \qquad b = \bar{R}_M - \bar{R}_Z$$

Putting this together and letting $\bar{R}_i$ and β_i be the expected return and beta on an asset or portfolio, the equation for the expected return on any security or portfolio becomes

$$\bar{R}_i = \bar{R}_Z + \left(\bar{R}_M - \bar{R}_Z\right)\beta_i \tag{14.1}$$

This is the so-called zero-beta version of the CAPM and is plotted in Figure 14.2. This form of the general equilibrium relationship is often referred to alternatively as a two-factor model.

Rigorous Derivation Assume for the moment that the market portfolio lies on the efficient frontier in expected return standard deviation space. Later in this chapter we show that it must, indeed, do so. In Chapter 6 we showed that the entire efficient frontier can be traced out by allowing the riskless rate of interest to vary and finding the tangency point

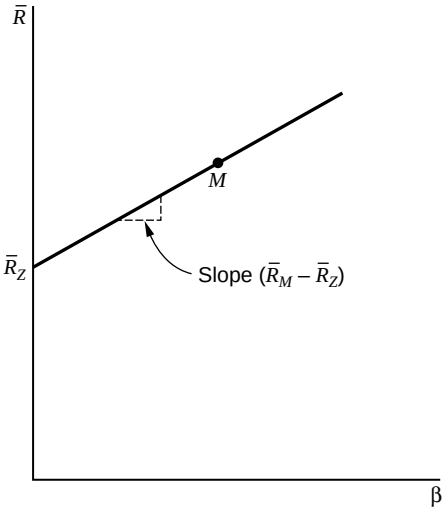


Figure 14.2 The zero-beta capital asset pricing line.

[4]To see that such a point exists, note that the straight line must go indefinitely in both directions. All positive combinations of *A* and *D* lie on the line segment between *A* and *D*. However, if we purchase *D* and sell *A* short, we move above *D*, and vice versa. Thus the line continues indefinitely and, in particular, cuts the vertical axis.

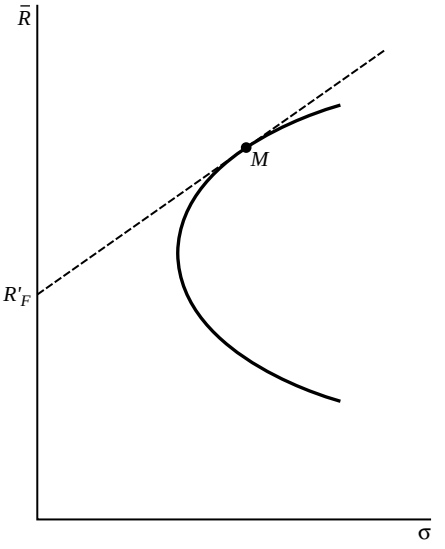


Figure 14.3 The opportunity set with rate R_F.

between the efficient frontier and a ray passing through the riskless rate (on the vertical axis). Corresponding to every "risk-free rate," there was one point on the efficient frontier, and vice versa. There is, of course, one unique riskless rate in the market (if any). Thus the procedure of varying the riskless rate was simply a method of obtaining the full efficient frontier. In all cases but one, what we called the riskless rate was an artificial construct we used to obtain one point on the efficient frontier. Define R'_F as the riskless rate such that if investors could lend and borrow unlimited amounts of funds at the rate R'_F, they would hold the market portfolio.

The investor who could lend and borrow at the riskless rate R'_F would face an investment opportunity set as depicted in Figure 14.3. To solve for optimal proportions, she would face a set of simultaneous equations directly analogous to Equation (13.4). One such equation is[5]

$$\lambda\left(X_1\sigma_{1j} + X_2\sigma_{2j} + \cdots + X_j\sigma_j^2 + \cdots + X_N\sigma_{Nj}\right) = \bar{R}_j - R'_F \tag{14.2}$$

Note that in the equation the X_is are market proportions because R'_F is defined as that value of the riskless rate that causes investors to hold the market portfolio.

In the previous chapter we showed that the term in parentheses in the left-hand side of Equation (14.2) was simply the covariance between the return on security j and the return on the market portfolio. Thus Equation (14.2) can be written as

$$\lambda \operatorname{cov}\left(R_j R_M\right) = \bar{R}_j - R'_F$$

or

$$\bar{R}_j = R'_F + \lambda \operatorname{cov}\left(R_j R_M\right) \tag{14.3}$$

[5]These equations are first-order conditions and must hold for the tangency point of any line drawn from the vertical axis and the efficient frontier.

The expected return on the market portfolio is a weighted average of the expected return on individual securities. Because Equation (14.3) holds for each security, it must also hold for the market. Thus

$$\bar{R}_M = R'_F + \lambda \operatorname{cov}(R_M R_M)$$

But $\operatorname{cov}(R_M R_M)$ is the variance of M so that

$$\bar{R}_M = R'_F + \lambda \sigma_M^2 \qquad \text{or} \qquad \lambda = \frac{\bar{R}_M - R'_F}{\sigma_M^2}$$

Substituting the expression for λ into Equation (14.3) and rearranging yields

$$\bar{R}_j = R'_F + \frac{\bar{R}_M - R'_F}{\sigma_M^2} \operatorname{cov}(R_j R_M)$$

or

$$\bar{R}_j = R'_F + \beta_j (\bar{R}_M - R'_F) \tag{14.4}$$

Note that a riskless asset with a return of R'_F does not really exist. However, there are an infinite number of assets and portfolios giving a return of R'_F. They are located along the solid portion of the line segment R'_F–C shown in Figure 14.4. Examine Equation (14.4). For R_j to be equal to R'_F, the last term must be zero. Thus any security or portfolio that has an expected return of R'_F must have a beta (covariance with the market portfolio) equal to zero.

Although equilibrium can be expressed in terms of any of the zero-beta portfolios on the solid portion of the line segment R'_F–C, it makes sense to utilize the least risky zero-beta portfolio. This is equivalent to the zero-beta portfolio that has the least total risk. We designate the minimum-variance zero-beta portfolio as Z and its expected return as $\bar{R}_Z$.

Then, because $\bar{R}_Z = R'_F$, the security market line can be written as

$$\bar{R}_j = \bar{R}_Z + (\bar{R}_M - \bar{R}_Z)\beta_j$$

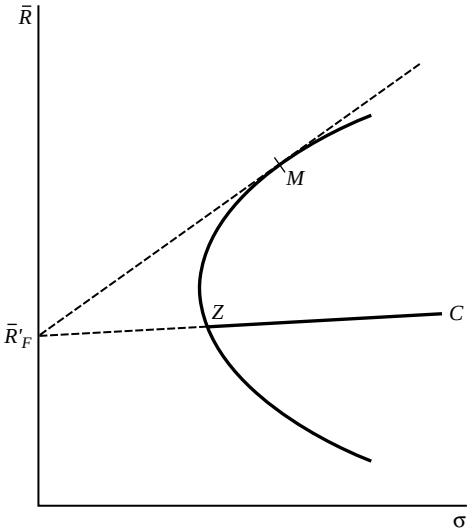


Figure 14.4 The location of portfolios with return R'_F.

This is exactly the expression [Equation (14.1)] we found for the security market line earlier in this chapter.

Let us see if we can learn anything about the location of this minimum-variance zero-beta portfolio. First, we know that the expected return on the zero-beta portfolio must be lower than the expected return on the market portfolio. The market portfolio is on the efficient segment of the minimum-variance frontier, and the slope at this point must be positive. Thus, as we move along the line tangent to $\bar{R}_M$ toward the vertical axis, we lower return. Because $\bar{R}_Z$ is the intercept of the tangency line and the vertical axis, it has a return less than $\bar{R}_M$. Second, as we prove later, the minimum-variance zero-beta portfolio cannot be efficient.

Proof Denote by *s* the portfolio that has the smallest possible variance. This portfolio can be formed as a combination of the market portfolio and the zero-beta portfolio:

$$\sigma_s^2 = X_Z^2\sigma_Z^2 + \left(1 - X_Z\right)^2\sigma_M^2$$

There is no covariance term because the covariance between these two assets is zero. To find the weights in each portfolio that minimize variance, take the derivative with respect to X_Z and set it equal to zero, or

$$\frac{d\sigma_s^2}{dX_Z} = 2X_Z\sigma_Z^2 - 2\sigma_M^2 + 2X_Z\sigma_M^2 = 0$$

Solving for X_Z,

$$X_Z = \frac{\sigma_M^2}{\sigma_M^2 + \sigma_Z^2}$$

Because both σ_M^2 and σ_Z^2 must be positive numbers, that portfolio with the smallest possible variance must involve positive weights on both the zero-beta and market portfolio. Because $\bar{R}_Z < \bar{R}_M$, portfolios of *Z* and *M* with positive weights must have higher expected returns than *Z*. Because *the* minimum-variance portfolio has higher return and smaller variance than *Z*, *Z* cannot be on the efficient portion of the minimum-variance frontier.

We can locate portfolios *Z*, *M*, and *s* on the minimum-variance frontier of all portfolios in expected return standard deviation space.[6] This is done in Figure 14.5. This figure presents the location of all efficient portfolios in expected return standard deviation space. All investors will hold some portfolio that lies along the efficient frontier (*SMC*). Investors who hold portfolios offering returns between *s* and $\bar{R}_M$ will hold combinations of the zero-beta portfolio and the market portfolio.[7] Investors who choose to hold portfolios to the right of *M* (choose returns above $\bar{R}_M$) will hold a portfolio constructed by selling portfolio *Z* short and buying the market portfolio. No investor will choose to hold only portfolio *Z*, for this is an inefficient portfolio. Furthermore, because investors in the aggregate hold the market portfolio, the aggregate holding of portfolio *Z* (long positions minus short positions) must be exactly zero. Note also that we still have a two mutual fund theorem. All investors can be satisfied by transactions in two mutual funds: the market portfolio and the minimum-variance zero-beta portfolio.

[6]The minimum-variance curve or minimum-variance frontier contains the set of portfolios that offers the lowest risk at any obtainable level of return. The efficient set (frontier) is a subset of these minimum-variance portfolios.

[7]Recall from Chapter 6 that the entire efficient frontier can be generated as portfolios of any two portfolios on the efficient frontier.

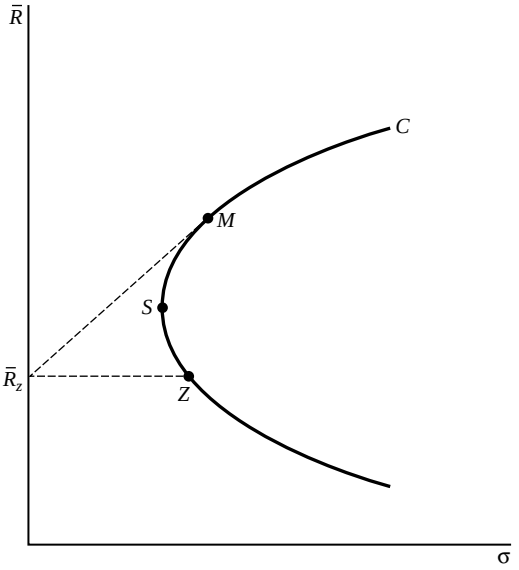


Figure 14.5 The minimum-variance frontier.

We started out this section by assuming that the market portfolio is efficient. Although we do not intend to provide a rigorous proof of its efficiency, a few comments should convince the reader of this truth. Those interested in a rigorous proof are referred to Fama (1970).

With homogeneous expectations, all investors face the same efficient frontier. Recall that with short sales allowed, all combinations of any two minimum variance portfolios are minimum variance. Thus, if we combine any two investors' portfolios, we have a minimum variance portfolio. The market portfolio is a weighted average or portfolio of each investor's portfolio where the weights are the proportion each investor owns of the total of all risky assets. Thus it is minimum variance. Because each investor's portfolio is efficient and return on the market is an average of the return on the portfolios of individual investors, the return on the market portfolio is the return of a portfolio on the efficient segment of the minimum-variance frontier. Thus the market portfolio is not only minimum variance but efficient.

Riskless Lending but No Riskless Borrowing

We have gone too far in changing our assumptions. As we agreed earlier, although it is unrealistic to assume that individuals can borrow at the riskless rate, it is realistic to assume that they can lend at a rate that is riskless. Individuals can place funds in government securities that have a maturity equal to their time horizon and, thus, be guaranteed of a riskless payoff at the horizon.

If we allow riskless lending, then the investor's choice can be pictured as in Figure 14.6.[8] As we argued in earlier chapters, all combinations of a riskless asset and a risky portfolio lie on the straight line connecting the asset and the portfolio. The preferred combination lies on the straight line passing through the risk-free asset and tangent to the efficient frontier. This is the line R_FT in Figure 14.6.

[8]Once again, we are assuming short sales are allowed. This is a necessary assumption.

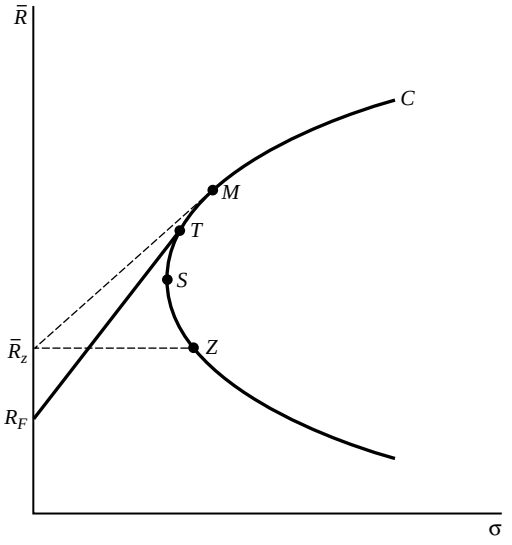


Figure 14.6 The opportunity set with riskless lending.

Notice that we have drawn *T* below and to the left of the market portfolio *M* and, hence, $\bar{R}_Z > R_F$. This was not an accident. Let us examine why this must hold. Before we introduced the ability to lend at the riskless rate, all investors held portfolios along the efficient frontier *SMC* (portfolios along the line $\bar{R}_Z M$ do not exist). With riskless lending, the investor can hold portfolios of riskless and risky assets along the line $R_F T$. If the investor chooses to hold an investment on the line $R_F T$, he would be placing some of his funds in the portfolio of risky assets denoted by *T* and some in the riskless asset. The choice to hold any portfolio of risky assets other than *T* would never be made. Now, why can't *T* and *M* be the same portfolio? As long as any investor has a risk-return trade-off such that she chooses to hold a portfolio of investments to the right of *T*, the market must lie to the right of *T*. For example, assume that all investors but one choose to lend money and hold portfolio *T*. Now this one investor who does not choose *T* must hold a portfolio to the right of *T* on the efficient frontier *STC*. If the investor did not, then he would be better off holding a portfolio on the line $R_F T$ and, hence, holding portfolio *T*. Because the market portfolio is an average of the portfolios held by all investors, the market portfolio must be a combination of the investor's portfolio and *T*. Thus it lies to the right of *T*. *M*, being to the right of *T*, leads directly to $\bar{R}_Z$ being larger than R_F. R_F is the intersection of the vertical axis and a line tangent to the efficient frontier at *T*. Similarly, $\bar{R}_Z$ is the intersection of the vertical axis and a line tangent at *M*. Because the slope of the efficient frontier at *M* is less than at *T* and because *M* lies above *T*, the line tangent at *M* must intersect the vertical axis above the line tangent at *T*.[9] Thus $\bar{R}_Z$ must be greater than R_F.

The efficient frontier is given by the straight line segment $R_F T$ and curve *TMC*.[10] Notice that, in the case of no lending and borrowing, combinations of all efficient portfolios were efficient. In the case where riskless lending is allowed, not all combinations of efficient

[9]The property of the two slopes follows directly from the concavity of the efficient frontier proved in Chapter 5.

[10]The reader might note that portfolio *T* is a corner portfolio, a portfolio whose composition is different from those immediately adjacent to it. All portfolios to the right of *T* on the efficient frontier are made up of combinations of portfolios *M* and *Z*, whereas those to the left of *T* are made up of portfolios *M* and *Z* plus the riskless security.

portfolios are efficient. It should be obvious to the reader that combinations of a portfolio from the line segment R_FT and a portfolio from the curve *TMC* are dominated by a portfolio lying along the curve *TMC*.

Portfolio *T* can be obtained by combining portfolios *Z* and *M*. Examining the efficient frontier, we see that investors who select a portfolio along the line segment R_FT are placing some of their money in portfolio *T* (which is constructed from the market portfolio plus portfolio *Z*) and some in the riskless asset. (Those who select a portfolio on the segment *TM* are placing some of their money in portfolio *M* and some in *Z*.) Those who select a portfolio on *MC* are selling portfolio *Z* short and investing all of the proceeds in *M*. (Notice that our two mutual fund theorem has been replaced with a three mutual fund theorem.) All investors can be satisfied by holding (long or short) some combination of the market portfolio, the minimum-variance zero-beta portfolio, and the riskless asset.[11]

Having examined all efficient portfolios in expected return standard deviation space, let us turn our attention to the location of securities and portfolios in expected return beta space. Let us develop the security market line.

The market portfolio *M* is still an efficient portfolio. Thus the analysis of the last section holds. All securities contained in *M* have an expected return given by

$$\bar{R}_j = \bar{R}_Z + \beta_j\left(\bar{R}_M - \bar{R}_Z\right) \tag{14.5}$$

Similarly, all portfolios composed solely of risky assets have their return given by Equation (14.5). This splits as a straight line in expected return beta space and is the line $\bar{R}_ZTMC$ in Figure 14.7. This equation holds only for risky assets and for portfolios of risky assets. It does not describe the return on the riskless asset or the return on portfolios that contain the riskless asset.

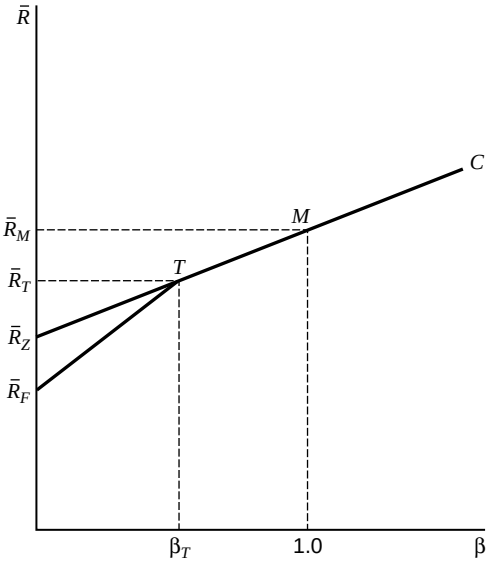


Figure 14.7 The location of investments in expected return beta space.

[11]Note that although we continually speak of using the market portfolio and the minimum-variance zero-beta portfolio to obtain the efficient frontier, any other two minimum-variance portfolios would serve equally well.

In the previous chapter we examined combinations of the riskless asset and a risky portfolio and found that they lie on the straight line connecting the two points in expected return beta space. Because investors who lend all hold risky portfolio *T*, the relevant line segment is R_FT in Figure 14.7.

Thus, although the straight line $\bar{R}_ZM$ can be thought of as the security market line for all risky assets and for all portfolios composed entirely of risky assets, it does not describe the return on portfolios (and, of particular note, on those efficient portfolios) that contain the riskless asset. Efficient portfolios have their return given by the two line segments R_FT and TC in Figure 14.7. The fact that efficient portfolios have lower return for a given level of beta than individual assets may seem startling. But remember that securities or portfolios on $\bar{R}_ZT$ have a higher standard deviation than portfolios with the same return on segment R_FT. (To understand this, remember that the return on portfolio *Z* is uncertain, even though it has a zero beta, whereas the return on the riskless asset is certain.)

Before moving on to other models, it is well worth reviewing certain characteristics of those we have been discussing, particularly insofar as they resemble or are different from the characteristics of the simple CAPM.

First, note that, under either of these models, all investors no longer hold the same portfolio in equilibrium. This is comforting, for it is more consistent with observed behavior. Of less comfort is that investors still hold most securities (either long or short) and hold many securities short. In the case where neither lending nor borrowing is allowed, we have a two mutual fund theorem. In the case where riskless lending is allowed, we have a three mutual fund theorem.

As in the case of the simple CAPM, we still get a security market line. In addition, many of the implications of this relationship are the same. For risky assets or portfolios, expected return is still a linearly increasing function of risk as measured by beta. It is only market risk that affects the return on individual risky securities and portfolios of risky securities. On these securities the investor gains no extra return from bearing diversifiable risk. In fact, the only difference lies in the intercept and slope of the security market line.[12]

Other Lending and Borrowing Assumptions

Brennan (1971) has analyzed the situation where riskless lending and borrowing is available, but at different rates. The efficient frontier for the individual when riskless borrowing and lending at different rates is possible was analyzed in Chapter 5. If all investors face the same efficient frontier, this efficient frontier must appear as in Figure 14.8.

In this diagram, *L* stands for the portfolio of risky securities that will be held by all investors who lend money, and *B* stands for the portfolio of all securities that will be held by investors who borrow money. The market portfolio must lie on the efficient frontier, and it must lie between *L* and *B*.

Let us examine why. The only portfolios of risky securities held by investors are *L* and *B* and intermediate portfolios on the curve *LB*. Earlier we showed that combinations of efficient portfolios were efficient. In the earlier section, lending and borrowing was not allowed so that the proof was that combinations of portfolios on the efficient portion of the minimum variance frontier were also on the efficient portion. The market portfolio is a weighted average of all portfolios held by individuals. Because these are efficient, we know from the earlier discussion that the market portfolio lies on the efficient portion of the minimum-variance curve. But we can be even more precise. The return on the market portfolio is a weighted

[12]In all models the efficient frontier itself is affected by diversifiable risk. Because the shape of the frontier affects the location of the tangency portfolio, diversifiable risk has some effect on security returns.

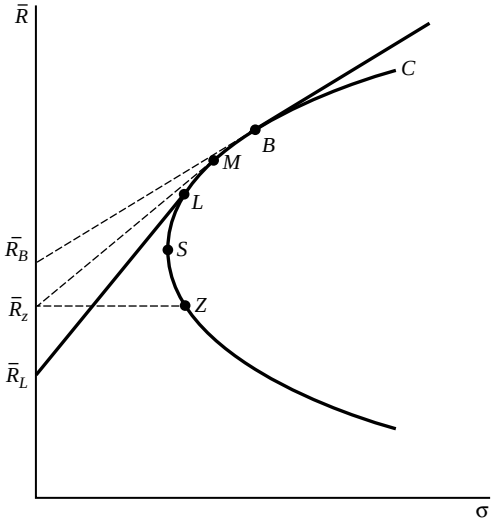


Figure 14.8 The opportunity set with a differential lending and borrowing rate.

average of the return of portfolio *L*, portfolio *B*, and all intermediate portfolios. Thus its return must be between *L* and *B*. Therefore the market portfolio must lie somewhere on the efficient frontier between *L* and *B*. Having established that the market portfolio lies on the efficient frontier between *L* and *B*, we derive, in the same manner, the same security market line as we derived in the last section of this chapter. Equation (14.1) still holds. However, remember that this equation only describes the return on securities and portfolios that do not have any investment in the riskless asset (long or short). Thus the equation will not describe the return on portfolios that are combinations of a risky portfolio and a riskless asset along the straight line between $\bar{R}_L$ and *L* or with return more than $\bar{R}_B$.

Brennan (1971) has also examined the case where the borrowing rate differed from the lending rate and where these rates were different for each investor. Once again, because the market portfolio lies on the efficient frontier, an equation identical in form to Equation (14.1) describes the return on all risky assets and on all portfolios composed entirely of risky assets.

PERSONAL TAXES

The simple form of the CAPM ignores the presence of taxes in arriving at an equilibrium solution. The implication of this assumption is that investors are indifferent between receiving income in the form of capital gains or dividends and that all investors hold the same portfolio of risky assets. If we recognize the existence of taxes and, in particular, the fact that capital gains are taxed, in general, at a lower rate than dividends, the equilibrium prices should change. Investors should judge the return and risk on their portfolio after taxes. This implies that, even with homogeneous expectations about the before-tax return on a portfolio, the relevant (after-tax) efficient frontier faced by each investor will be different. However, a general equilibrium relationship should still exist because, in the aggregate, markets must clear. In the appendix at the end of this chapter we derive the general equilibrium pricing equation for all assets and portfolios, given differential taxes on income and capital gains. The return on any asset or portfolio is given by

$$E(R_i) = R_F + \beta_i\left[\left(E(R_M) - R_F\right) - \tau(\delta_M - R_F)\right] + \tau(\delta_i - R_F) \tag{14.6}$$

where

δ_M = the dividend yield (dividends divided by price) of the market portfolio

δ_i = the dividend yield for stock i

τ = a tax factor that measures the relevant market tax rates on capital gains and income. τ is a complex function of investors' tax rates and wealth. However, it should be a positive number. See the appendix for further discussion.

The equilibrium relationship for expected returns has now become very complex. When dividends are on average taxed at a higher rate than capital gains (as they are in the U.S. economy), τ is positive, and expected return is an increasing function of dividend yield. This is intuitively appealing because the larger the fraction of return paid in the form of dividends, the more taxes the investor will have to pay and the larger the pretax return required. The reader may wonder why the last term contains R_F as well as the dividend yield. The reason for this is the tax treatment of interest on lending and borrowing. Because interest payments are for all intents and purposes taxed at the same rate as dividends, they enter the relationship in a parallel manner, although with an opposite sign.[13] The fact that the term in square brackets has the correct form can be seen by letting security i be the market portfolio and noting that (because beta equals 1 for the market portfolio) the equation reduces to $E(R_M) = E(R_M)$.

Examination of Equation (14.6) reveals that a security market line is no longer sufficient to describe the equilibrium relationship. In previous versions of general equilibrium relationships, the only variable associated with the individual security that affected expected return was its beta. Now we see from Equation (14.6) that both the securities beta and its dividend yield affect expected return. This means that equilibrium must be described in three-dimensional space (R_i, β_i, δ_i) rather than two-dimensional space. The resultant equilibrium relationship [Equation (14.6)] will be a plane rather than a straight line. The plane will be located such that for any value of beta, expected return goes up as dividend yield goes up, and for any value of dividend yield, expected return goes up as beta goes up. We will have more to say about the location of the plane (the parametrization of this equation) in the next chapter.

If returns are determined by an equilibrium model like that presented in Equation (14.6), it should be possible to derive optimal portfolios for any investor as a function of the tax rates paid on capital gains and dividends. Although the mathematics of the solution are rather complex, the economic intuition behind the results is strong.[14] All investors will hold widely diversified portfolios that resemble the market portfolio, except they will be tilted in favor of those stocks in which the investor has a competitive advantage. For example, investors whose tax bracket is below the average effective rate in the market should hold more of high-dividend stocks in their portfolio than the percentage these stocks constitute of the market portfolio, while they should hold less (and in extreme cases even short sell) stocks with very low dividends. Low-tax-bracket investors have a comparative advantage in holding high-dividend stocks for the tax disadvantage of these stocks is less disadvantageous to them than it is to the average stockholder. Individual

[13]The implications of this for investor behavior are interesting. For example, an investor could convert a dividend-paying stock into one with only a capital gains return by borrowing a sum of money such that when the sum borrowed plus the initial planned investment in a stock is invested in the stock, interest payments exactly equal the dividend payments on the stock.

[14]See Elton and Gruber (1978) for the derivation of the composition of optimal portfolios under taxation.

investors in the market seem to behave as the analysis suggests they should.[15] The optimization rules described in Elton and Gruber (1978) ensure that markets will clear at the returns established in Equation (14.6).

NONMARKETABLE ASSETS

Up to now, we have assumed that all assets are readily marketable so that each investor was free to adjust her portfolio to an optimum. In truth, every investor has nonmarketable assets, or assets that she will not consider marketing. Human capital is an example of a nonmarketable asset. People are forbidden by law from selling themselves into slavery in the United States. There is no direct way that an investor can market her claims to future labor income. Similarly, the investor has other future monetary claims, such as social security payments or the future payments from a private retirement program, that cannot be marketed. There are categories of marketable assets that, although the investor might be able to market them, he considers a fixed part of the portfolio. For example, an investor who owns his own home can market it, but he will often not consider switching houses as part of changes in his "optimum investment portfolio." This is due, in part, to large transaction costs but also to nonmonetary factors.

If we divide the world up into marketable and nonmarketable assets, then a simple equation exists for the equilibrium return on all assets. Let

R_H equal the one-period rate of return on nonmarketable assets

P_H equal the total value of all nonmarketable assets

P_M equal the total value of all marketable assets

All other terms are defined as before. Then, it can be shown that[16]

$$E(R_j) = R_F + \frac{E(R_M) - R_F}{\sigma_M^2 + P_H/P_M \operatorname{cov}(R_M R_H)}\left[\operatorname{cov}(R_j R_M) + \frac{P_H}{P_M}\operatorname{cov}(R_j R_H)\right]$$

To contrast this with the simple CAPM, we can write the simple model as

$$E(R_j) = R_F + \frac{E(R_M) - R_F}{\sigma_M^2}\left[\operatorname{cov}(R_j R_M)\right]$$

Notice that the inclusion of nonmarketable assets leads to a general equilibrium relationship of the same form as the simple model that excluded nonmarketable assets. However, the market trade-off between return and risk is different, as is the measure of risk for any asset. Including nonmarketable assets, the market risk–return trade-off becomes

$$\frac{E(R_M) - R_F}{\sigma_M^2 + \frac{P_H}{P_M}\operatorname{cov}(R_M R_H)}$$

[15]Pettit and Stanley (1979) have found that investors tend to behave as this model suggests they should behave.

[16]For a derivation, see Mayers (1972). Although this equation does not appear in Mayers, it can be derived from his Equation (19) with a little algebra. The reader may be bothered by the fact that P_H appears in our equation whereas the Mayers's equations make use of the income (actually income plus value) of the asset one period hence. However, there is no inconsistency, as Mayers's Equation (15) allows for the determination of P_H.

rather than

$$\frac{E(R_M) - R_F}{\sigma_M^2}$$

It seems reasonable to assume that the return on the total of nonmarketable assets is positively correlated with the return on the market, which would suggest that the market return-risk trade-off is lower than that suggested by the simple form of the model. How much lower is a function of both the covariance between the return on the nonmarketable assets and the marketable assets and the total value of nonmarketable assets relative to marketable assets. If nonmarketable assets had a very small value relative to marketable assets or if there was an extremely low correlation between the return on marketable and nonmarketable assets, there would be little harm done in using the standard CAPM. However, it seems likely that because nonmarketable assets include, at a minimum, human capital, and because wage rates as well as market performance are correlated with the performance of the economy, there will be important differences between these models.

In addition, the definition of the risk of any asset has been changed. With nonmarketable assets, it is a function of the covariance of an asset with the total stock of nonmarketable assets, as well as with the total stock of marketable assets. The weight this additional term receives in determining risk depends on the total size of nonmarketable assets relative to marketable assets. The risk on any asset that is positively correlated with the total of nonmarketable assets will be higher than the risk implied by the simple form of the CAPM.

Considering the difference in both the reward–risk ratio and the size of risk itself, we can see that the equilibrium return for an asset can be either higher or lower than it is under the standard form of the CAPM. If the asset is negatively correlated with the total of nonmarketable assets, its equilibrium return will be lower for its risk and the price of risk will be lower. However, if its return is positively correlated with the return on marketable assets, its equilibrium return could be higher or lower, depending on whether the increased risk is high enough to offset the decreased market price of risk.

Mayers (1972) explores the implications of his model for the optimal portfolio holdings of individuals. As you would suspect, investors tilt their portfolios, holding a smaller percentage of those stocks (than found in the market) with which their nonmarketable securities are most highly correlated.

Brito (1977, 1978) has examined, in more detail, the optimum portfolio holdings of individuals in equilibrium when nonmarketable assets are present. He finds that each individual can select an optimal portfolio from among three mutual funds. The first mutual fund is a portfolio that has a covariance with each marketable asset equal in magnitude but opposite in sign to the covariance between the investor's nonmarketable portfolio and each marketable asset. Note two things about this fund: first, it will have a different composition for different investors, according to the nonmarketable assets they hold; second, the reason for its optimality has an intuitive explanation—it is that portfolio that diversifies away as much of the nonmarketable risk as it is possible to diversify away. In short, it allows the investor to "market" as much of her nonmarketable assets as is possible. Brito then shows that each individual will allocate the remainder of her wealth between the riskless security (the second fund) and a third fund that is the market portfolio minus the *aggregate* of all investments made in the first type of fund by all investors. Note that, while the second and third funds are the same for all investors, the first fund has a different composition for each investor, according to the composition of her nonmarketable assets.

At the same time as Mayers's analysis is important for the insight it provides into the pricing of nonmarketable assets, it is at least as important for the insight it gives us into the

missing asset problem. Empirical tests of general equilibrium models will always have to be conducted with the market defined as including something less than the full set of assets in the economy. The equilibrium equations described previously are perfectly valid for examining the missing asset problem, where R_M is now defined as the return on the collection of assets selected to represent the market and R_H is the return on the assets that were left out. In a manner exactly parallel to that presented, they allow us to think through the influence of missing assets on both the market's risk-return trade-off and the equilibrium return from missing assets.

HETEROGENEOUS EXPECTATIONS

Several researchers have examined the existence and characteristics of a general equilibrium solution when investors have heterogeneous expectations.[17] Although all of these models lead to forms of an equilibrium pricing equation that have some similarity to those presented earlier in this chapter and in the last chapter, there are important differences. Equilibrium can still be expressed in terms of expected returns, covariances, and variances, but now these returns, covariances, and variances are complex weighted averages of the estimates held by different individuals. The weightings are very complex because they involve information about investor utility functions. In particular, they involve information about investors' trade-offs (marginal rate of substitution) between expected return and variance. But this trade-off for most utility functions is a function of wealth and, hence, prices. This means that prices are required to determine the risk–return trade-offs that we need to determine prices. Thus, in general, an explicit solution to the heterogeneous expectation problem cannot be reached. The problem can be made simpler by placing additional restrictions either on investor utility functions or on the characteristics of opportunities facing the investor.

The first approach was taken by Lintner (1969). He could not derive a simple CAPM under heterogeneous expectations because the marginal rate of substitution between expected return and variances was, itself, a function of equilibrium prices. If we assume a utility function such that the marginal rate of substitution is not a function of wealth, then we will not face this problem. We have already studied such a class of utility functions in Chapter 10. They were the functions exhibiting constant absolute risk aversion. Lintner assumed this type of function (to be precise, he assumed a negative exponential utility function).[18] Utilizing this function, he showed that the Sharpe–Lintner–Mossin form of the CAPM model holds and that the term $(\bar{R}_M - R_F)/\sigma_M^2$ in Equation (13.2) is proportional to the harmonic mean of the risk-avoidance coefficient, and all expected values, variances, and covariances are complicated averages of the probability beliefs and risk preferences of all individuals.

A second way to arrive at more testable models of equilibrium under heterogeneous assumptions is to place restrictions on the form that the heterogeneity can assume. Gonedes (1976) assumes that a set of basic economic activities exists such that any firm can be viewed as some combination of these basic economic activities and the heterogeneous expectations arise because of disagreement about the exact combination (weighting) of those basic economic activities that represent a firm. Gonedes analyzes the case where this is the one source of heterogeneous expectations. He shows that, under this assumption, the minimum-variance frontier is the same for all investors, even though they have heterogeneous expectations about

[17]See Lintner (1969), Sharpe (1970), Fama (1976), and Gonedes (1976).

[18]Lintner assumes the negative exponential utility function given by $u(w) = e^{-a_i W_i}$. The measure of risk aversion is given by a_i.

the returns from different securities. Furthermore, the market portfolio is a minimum-variance portfolio for each and every investor. Gonedes then proceeds to show that beta is a sufficient measure of risk and that the equilibrium models lead to a linear relationship between expected return and beta parallel to that found under simpler forms of the CAPM.

NON-PRICE-TAKING BEHAVIOR

Up to now we have assumed that individuals act as price takers in that they ignore the impact of their buying or selling behavior on the equilibrium price of securities and, hence, on their optimal portfolio holdings. The obvious question to ask is what happens if there are one or more investors, such as mutual funds or large pension funds, who believe that their behavior impacts price. The method of analysis used by Lindenberg (1976, 1979) derives equilibrium conditions under all possible demands by the price affector. The price affector selects her portfolio to maximize utility given the equilibrium prices that will result from her action. Assuming that the price affector operates so as to maximize utility, we can then arrive at equilibrium conditions. Lindenberg finds that all investors, including the price taker, hold some combination of the market portfolio and the riskless asset. However, the price affector will hold less of the riskless asset (will be less of a risk avoider) than would be the case if the price affector did not recognize the fact that her actions affected price. By doing so, the price affector increases utility. Because the price affector still holds a combination of the riskless asset and the market portfolio, we still get the simple form of the CAPM, but the market price of risk is lower than it would be if all investors were price takers.

Lindenberg (1979) goes on to analyze collective portfolio selection and efficient allocation among groups of investors. He finds that by colluding or merging, individuals or institutions can increase their utility. This analysis provides us with one reason for the existence of large financial institutions.

MULTIPERIOD CAPM

Up to now, we have assumed that all investors make investment decisions based on a single-period horizon. In fact, the portfolio an investor selects, at any point in time, is really one step in a series of portfolios that he intends to hold over time to maximize his utility of lifetime consumption. Two questions immediately become apparent:

1. What are the conditions under which the simple CAPM adequately describes market equilibrium?
2. Is there a fully general multiperiod equilibrium model?

Fama (1970) and Elton and Gruber (1974, 1975) have explored the conditions under which the multiperiod investment consumption decision can be reduced to the problem of maximizing a one-period utility function. These conditions are as follows:

1. The consumer's tastes for particular consumption goods and services are independent of future events (any future sets of conditions).
2. The consumer acts as if consumption opportunities in terms of goods and their prices are known at the beginning of the decision period (are not state dependent).[19]
3. The consumer acts as if the distribution of one-period returns on all assets are known at the beginning of the decision period (are not state dependent).

[19]A process is not state dependent if its outcomes do not depend on which one of a set of events occurs.

Furthermore, Fama (1970) has shown that if the investor's multiperiod utility function, expressed in terms of multiperiod consumption, exhibits both a preference of more to less and risk aversion with respect to each period's consumption, then the derived one-period utility has the same properties with respect to that period's consumption.

Recall earlier that risk aversion and preferring more to less were two assumptions necessary to obtain an efficient frontier. If we make the additional assumptions of the standard CAPM, we obtain the standard CAPM even for investors with a multiperiod horizon. If we make the additional assumptions underlying the zero-beta version of the CAPM, the zero-beta model is appropriate for investors with a multiperiod horizon. In short, the Fama multiperiod assumptions make single-period capital asset pricing models appropriate for investors with multiperiod horizons. The particular single-period model that results depends on the additional assumptions that are being made.

THE MULTI-BETA CAPM

Merton (1973) has constructed a generalized intertemporal CAPM in which a number of sources of uncertainty would be priced. Merton models investors as solving lifetime consumption decisions when faced with multiple sources of uncertainty. In this multiperiod setting, uncertainty exists not only about the future value of securities but also about such other influences as future labor income, future prices of consumption goods, future investment opportunities, and so on. Investors will form portfolios to hedge away each of these risks (to the extent possible). If sources of risk are a general concern to investors, then these sources of risk will affect the expected returns on securities. The inflation model is the simplest form of a multi-beta CAPM where the expected return on any security can be expressed as a function of two sensitivities,

$$\overline{R}_i - R_F = \beta_{iM}\left(\overline{R}_M - R_F\right) + \beta_{iI}\left(\overline{R}_I - R_F\right)$$

This expression represents the standard CAPM plus a new term. The new term is the product of a new beta (which is the sensitivity of any security to the portfolio of securities that is held to hedge away inflation risk) and the price of inflation risk.

The multi-beta CAPM tells us that the expected return on any security should be related to the security's sensitivity to a set of influences. The form of the expected return is

$$\overline{R}_i - R_F = \beta_{iM}\left(\overline{R}_M - R_F\right) + \beta_{iI1}\left(\overline{R}_{I1} - R_F\right) + \beta_{iI2}\left(\overline{R}_{I2} - R_F\right) + \cdots$$

In this relationship, all of the $\overline{R}_{Ij}$s are expected returns on a set of portfolios that allows the investor to hedge a set of risks with which he or she is concerned. Although the theory tells us that these should be additional influences present in pricing securities and that these influences should be related to the investor's multiperiod utility functions, it does not tell us explicitly what these influences are or exactly how to form portfolios to hedge whatever risks they represent. One set of risks we might consider as potentially important is the four risks (in addition to the market) that we examined in Chapter 8: default risk, term structure risk, deflation risk, and profit risk.

We leave this subject at this point but return to it in a later chapter, when we discuss arbitrage pricing theory.

CONSUMPTION CAPM

John Cochrane's (2001) textbook offers a compelling paradigm for asset pricing, starting with a consumption-based intertemporal equilibrium model, bypassing the normal development via single-period portfolio theory that leads to the traditional CAPM. This is a very

elegant and persuasive treatment that generalizes CAPM to a multiperiod economy and has direct implications for derivative pricing. Finally, and perhaps most importantly, well-trained economists find this approach provides an intuitive and persuasive access to the central results of the financial economics literature.

This approach builds on earlier work by Breeden (1979) and Rubinstein (1976). The investor's problem is to allocate wealth to maximize the utility of consuming both now and in the future. In other words, the investor is faced with an intertemporal choice problem of the form

$$\text{Max } E_t\left[\sum_{j=0}^{\infty}\delta^j U(c_{t+j})\right]$$

where c_{t+j} represents future consumption and δ is a subjective discount factor applied to future consumption. For a given budget constraint, the first-order conditions for this problem imply that

$$U'(c_t) = \delta^j E_t\left[(1+R_{i,t+j})U'(c_{t+j})\right]$$

for all assets i and periods j into the future.[20] Dividing through by the marginal utility of consumption today, we have the important result that

$$1 = E_t\left[(1+R_{i,t+j})m_{t,j}\right]$$

where $m_{t,j} = \delta^j \frac{U'(c_{t+j})}{U'(c_t)}$ is the intertemporal marginal rate of substitution. It also has the interesting interpretation of being a *stochastic discount factor* (sometimes also referred to as a *pricing kernel*) because it takes an asset with uncertain per dollar future payoff $\$(1 + R_{i,t+j})$ back to the present to be valued at \$1. If there is a riskless asset in this economy with return $R_{F,t+j}$, then $1 = E_t[(1 + R_{F,t+j})m_{t,j}] = (1 + R_{F,t+j})E_t[m_{t,j}]$ or $E_t[m_{t,j}] = \frac{1}{(1+R_{F,t+j})}$, so that the expected value of the stochastic discount factor is equal to the discount factor used when the future payment is in fact without any risk. For the subsequent discussion we will drop the time subscripts.

Cochrane (2001) argues that this intertemporal equilibrium model provides a straightforward way to value all financial claims. Security prices are determined so that the expected value of the growth of a dollar invested discounted by the stochastic discount factor equals the value of a dollar today. This is a direct implication of the equation $1 = E_t[(1 + R_{i,t+j})m_{t,j}]$. This gives rise immediately to a beta pricing model[21] $\bar{R}_i = R_F + \beta_{i,m}\lambda_m$, where $\beta_{i,m} = \frac{\sigma_{im}}{\sigma_m^2}$ is the beta of the security on this stochastic discount factor m and $\lambda_m = -\frac{\sigma_m^2}{\bar{m}}$.

The difficulty is, however, that the stochastic discount factor m is not observable. There are three general approaches to this problem. The first is to specify m directly through

---

[20]Cochrane (2001) along with many economists defines return as the growth in value of \$1 invested, or ending value; we have changed the notation to be consistant with the rest of the book, where ending value in period $t + j$ of a dollar invested in period t is $\$(1 + R_{t,t+j})$ where $R_{t,t+j}$ is the holding period return including income and capital gains from period t to period $t + j$.

[21]From the result (suppressing time subscripts) $1 = E[(1 + R_i)m] = E(1 + R_i)\ \bar{m} + \sigma_{im}$, substituting $\bar{m} = 1/(1 + R_F)$, we have $\bar{R}_i = R_F - \frac{\sigma_{im}}{\bar{m}} = R_F + \beta_{i,m}\lambda_m$, where $\beta_{i,m} = \frac{\sigma_{im}}{\sigma_m^2}$ and $\lambda_m = -\frac{\sigma_m^2}{\bar{m}}$.

assumptions made about utility and using measures of consumption. Suppose, for example, that the utility for consumption can be adequately represented by a power utility function of the form $U(c_t) = \frac{1}{1-\gamma} c_t^{1-\gamma}$. Then the stochastic discount factor $m = \delta(1 + C)^{-\gamma}$, where C is the growth rate of consumption. Substituting this expression into the preceding beta pricing model, to a first-order linear approximation, we have the Consumption beta asset pricing model of Breeden (1979):

$$\overline{R}_i = R_F + \beta_{it}.\lambda_t$$

The chief difficulty associated with this approach is obtaining accurate and timely measures of aggregate consumption *c*, which explains in part the poor empirical performance of this model (Hansen and Singleton 1982; Breeden, Gibbons, and Litzenberger 1989). An alternative approach is to use a vector of factors, some combination of which can proxy for consumption growth. Lettau and Ludvigson (2001) find that scaling the consumption growth factor by a lagged consumption–wealth ratio (CAY) leads to a considerable improvement in the empirical performance of the Consumption CAPM. Li, Vassalou, and Xing (2006) use investment growth rates for households, nonfinancial corporations, and the noncorporate sector, while Yogo (2006) incorporates durable consumption by including market return and durable and nondurable consumption. These models could be thought of as variants of the arbitrage pricing theory (APT) considered in Chapter 16. A third idea originally from Hansen and Jagannathan (1991) is that because the stochastic discount factor prices all financial claims, we might be just as well off inferring the stochastic discount factor from the observed set of asset returns. This important insight allows us to interpret the stochastic discount factor in terms of the mean–variance efficient portfolio, which, as we show in the appendix to Chapter 13, provides an interpretation that yields the standard CAPM as a direct implication.

CONCLUSION

In this chapter we have shown that the simple form of the CAPM is remarkably robust. Modifying some of its assumptions leaves the general model unchanged, whereas modifying other assumptions leads to the appearance of new terms in the equilibrium relationship or, in some cases, to the modification of old terms. That the CAPM changes with changes in the assumptions is not unusual. What is unusual is (1) the robustness of the methodology, in that it allows us to incorporate these changes, and (2) the fact that many of the conclusions of the original model hold, even with changes in assumptions.

The reader should be warned, however, that these results may seem stronger than they are. We have modified the assumptions one at a time. When assumptions are modified simultaneously, the departure from the standard CAPM may be much more serious. For example, when short sales were disallowed but lending and borrowing were allowed, the standard CAPM held. When riskless lending and borrowing were disallowed but short sales were allowed, we got a model that very much resembled the standard CAPM, except that the slope and intercept were changed. Ross (1977) has shown that when both riskless lending and borrowing and short sales are disallowed, one cannot derive a simple general equilibrium relationship.

There is no doubt that the general equilibrium models we now have are imperfect. The question is how well they describe conditions in the capital markets. We turn to this subject in the next chapter.

APPENDIX

DERIVATION OF THE GENERAL EQUILIBRIUM WITH TAXES

Earlier in this chapter we saw that any security or portfolio has an equilibrium return given by

$$\overline{R}_j = \overline{R}_Z + \left(\overline{R}_M - \overline{R}_Z\right)\frac{\sigma_{jM}}{\sigma_M^2}$$

We derived this expression by maximizing

$$\theta = \frac{\overline{R}_P - R'_F}{\sigma_P}$$

for the investor's portfolio (P) equal to the market portfolio M and the riskless rate defined as the intercept of a line tangent to point M. $\overline{R}_Z$ in the foregoing solution is the return on the minimum-variance portfolio that is uncorrelated with the portfolio M.

We could have repeated this analysis for any portfolio P different from M, and for assets included in portfolio P, we would get the following equilibrium relationship:

$$\overline{R}_j = \overline{R}_{0P} + \left(\overline{R}_P - \overline{R}_{0P}\right)\frac{\sigma_{jP}}{\sigma_P^2}$$

where $\overline{R}_{0P}$ is the expected return on the minimum-variance portfolio that is uncorrelated with portfolio P.

We will now make several changes in this expression. In a world of taxes, investors will reach equilibrium in terms of after-tax returns. The superscript A will be added to each variable to show that it holds in after-tax terms. In addition, the portfolio selected by each investor may be different because homogeneous before-tax expectations will produce heterogeneous after-tax expectations. Thus we will use the subscript i to stand for investor i. Finally, because we are assuming unlimited lending and borrowing, an asset exists (the riskless asset) that is uncorrelated with all portfolios. Thus we can replace $\overline{R}_{0P}$ with R_F. With these changes, the preceding equation can be written as

$$\overline{R}_{ji}^A = R_{Fi}^A + \left(\overline{R}_{Pi}^A - R_{Fi}^A\right)\frac{\text{cov}\left(R_{ji}^A R_{Pi}^A\right)}{\left(\sigma_{Pi}^A\right)^2} \tag{A.1}$$

While expectations of after-tax returns are heterogeneous, expectations of before-tax returns are homogeneous. We can write this expression in terms of before-tax returns.

Let

δ_j = the dividend yield on stock j

t_{di} = stockholder i's marginal tax rate on interest and dividends

t_{gi} = stockholder i's marginal tax rate on capital gains

w_i = the amount of stockholder i's wealth invested in risky assets

W = the sum of all wealth invested in risky assets

$$W = \sum_i w_i$$

Then,

$$\begin{aligned}\overline{R}_{ji}^{A} &= \left(\overline{R}_j - \delta_j\right)\left(1 - t_{gi}\right) + \delta_j\left(1 - t_{di}\right) \\ &= \overline{R}_j\left(1 - t_{gi}\right) - \delta_j\left(t_{di} - t_{gi}\right) \\ R_{Fi}^{A} &= R_F\left(1 - t_{di}\right)\end{aligned}$$

If we assume that next period's dividend is sufficiently predictable, then we can treat it as a certain stream and

$$\begin{aligned}\operatorname{cov}\left(R_{ji}^{A} R_{Pi}^{A}\right) &= \operatorname{cov}\left(R_j R_{Pi}\right)\left(1 - t_{gi}\right)^2 \\ \left(\sigma_{Pi}^{A}\right)^2 &= \sigma_{Pi}^2\left(1 - t_{gi}\right)^2\end{aligned}$$

Substituting in Equation (A.1),

$$\left(\overline{R}_j - R_F\right)\left(1 - t_{gi}\right) - \left(\delta_j - R_F\right)\left(t_{di} - t_{gi}\right) = \frac{\overline{R}_{Pi}^{A} - R_{Fi}^{A}}{\sigma_{Pi}^2}\operatorname{cov}\left(R_j R_{Pi}\right)$$

Dividing through by $1 - t_{gi}$, and multiplying through by w_i, and dividing through by λ_i, where λ_i is defined as

$$\frac{\overline{R}_{Pi}^{A} - R_{Fi}^{A}}{\sigma_{Pi}^2}\frac{1}{\left(1 - t_{gi}\right)}$$

we get

$$\frac{w_i}{\lambda_i}\left(\overline{R}_j - R_F\right) - \left(\delta_j - R_F\right)\frac{\left(t_{di} - t_{gi}\right)}{\left(1 - t_{gi}\right)}\frac{w_i}{\lambda_i} = w_i \operatorname{cov}\left(R_j R_{Pi}\right) \tag{A.2}$$

Summing this equation across all investors and dividing by Σw_i,

$$\left(\overline{R}_j - R_F\right)\frac{\sum_i \left(w_i/\lambda_i\right)}{\sum_i w_i} - \left(\delta_j - R_F\right) \times \left[\sum_i \frac{\left(t_{di} - t_{gi}\right)w_i}{\left(1 - t_{gi}\right)\lambda_i} \Big/ \sum_i w_i\right] = \frac{\sum_i w_i \operatorname{cov}\left(R_j R_{Pi}\right)}{\sum_i w_i}$$

But note that because

$$\frac{\sum_i w_i R_{Pi}}{\sum_i w_i} = R_M$$

the right-hand side of this equation is equal to $\text{cov}(R_j R_M)$. Define the following symbols:

$$H = \left(\sum_i w_i\right) \Big/ \left(\sum_i w_i/\lambda_i\right)$$

$$\tau = H\left(\sum_i \frac{(t_{di} - t_{gi})w_i}{(1 - t_{gi})\lambda_i}\right) \Big/ \sum_i w_i$$

We can see that the tax factor τ is a complex weighted average of the investor's tax rates, where the weights on each investor's tax rate is a function of the wealth he places in risky securities and his degree of risk avoidance as expressed by the ratio of excess return to variance on the portfolio he chooses to hold. Equation (A.2) can now be written as

$$(\bar{R}_j - R_F) - (\delta_j - R_F)\tau = H \,\text{cov}(R_j R_M) \tag{A.3}$$

Because expression (A.3) must hold for any asset or portfolio, it must hold for the market portfolio. Thus

$$(\bar{R}_M - R_F) - (\delta_M - R_F)\tau = H\sigma_M^2$$

or

$$H = \frac{(\bar{R}_M - R_F) - (\delta_M - R_F)\tau}{\sigma_M^2}$$

Substituting the expression for H into the equation and rearranging yields

$$\bar{R}_j = R_F + \frac{(\bar{R}_M - R_F) - (\delta_M - R_F)\tau}{\sigma_M^2}\,\text{cov}(R_j R_M) + (\delta_j - R_F)\tau$$

or

$$\bar{R}_j = R_F + \beta_j\left[(\bar{R}_M - R_F) - (\delta_M - R_F)\tau\right] + (\delta_j - R_F)\tau$$

QUESTIONS AND PROBLEMS

1. Assume the equilibrium equation shown below. What is the return on the zero-beta portfolio and the return on the market assuming the zero-beta model holds?

$$\bar{R}_i = 0.04 + 0.10\beta_i$$

2. In the previous chapter we showed that the standard CAPM model could be written in price form. What is the zero-beta model in price form?

3. Given the model shown below, what is the risk-free rate if the posttax equilibrium model describes returns?

$$\bar{R}_i = 0.05 + 0.10\beta_i + 0.24\delta_i$$

4. Given the following situation:

$$\bar{R}_M = 15 \quad \sigma_M = 22$$
$$\bar{R}_Z = 5 \quad \sigma_Z = 8$$
$$R_F = 3$$

draw the minimum variance curve and efficient frontier in expected return standard deviation space. Be sure to give the coordinates of all key points. Draw the security market line.

5. You have just lectured two tax-free institutions on the necessity of including taxes in the general equilibrium relationship. One believed you and one did not. Demonstrate that if the model holds, the one that did could engage in risk-free arbitrage with the one that did not in a manner such that:

A. Both parties believed they were making an arbitrage profit in the transaction.

B. The one who believed in the posttax model actually made a profit; the other institution incurred a loss.

6. Assume that returns are generated as follows:

$$R_i = \bar{R}_i + a_i\left(R_M - \bar{R}_M\right) + b_i(C - C)$$

where C is the rate of change in interest rates. Derive a general equilibrium relationship for security returns.

7. If $\bar{R}_M = 15\%$ and $R_F = 5\%$ and risk-free lending is allowed but riskless borrowing is not, sketch what the efficient frontier might look like in expected return standard deviation space. Sketch the security market line and the location of all portfolios in expected return beta space. Label all points and explain why you have drawn them as you have.

8. Assume you paid a higher tax on income than on capital gains. Furthermore, assume that you believed that prices were determined by the posttax CAPM. Now another investor comes along who believes that prices are determined by the pretax CAPM. Demonstrate that you can make an excess return by engaging in a two-security swap with him.

9. As we will see in the next chapter, most tests of the CAPM involve tests on common stock data and perform the tests using the S&P index. You have just had a revelation that bonds are also marketable assets and thus should belong in the market return. Show what effect leaving them out might have on stocks with different characteristics.

BIBLIOGRAPHY

1. Alexander, Gordon. "An Algorithmic Approach to Deriving the Minimum-Variance Zero-Beta Portfolio," *Journal of Financial Economics*, **4,** No. 2 (March 1977), pp. 231–236.
2. Arzac, Enrique, and Bawa, Vijay. "Portfolio Choice and Equilibrium in Capital Markets with Safety-First Investors," *Journal of Financial Economics*, **4,** No. 3 (May 1977), pp. 277–288.
3. Black, Fischer. "Capital Market Equilibrium with Restricted Borrowing," *Journal of Business*, **45,** No. 3 (July 1972), pp. 444–455.
4. Borch, Karl. "Equilibrium, Optimum and Prejustices in Capital Markets," *Journal of Financial and Quantitative Analysis*, **IV,** No. 1 (March 1969), pp. 4–14.
5. Breeden, D. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, **7** (1979), pp. 265–296.

6. ——. "Consumption Risk in Futures Markets," *Journal of Finance*, **35** (1980), pp. 503–520.
7. Breeden, D., and Litzenberger, R. "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, **51** (1978), pp. 621–651.
8. Breeden, D., Gibbons, M., and Litzenberger, R. "Empirical Tests of the Consumption-Oriented CAPM," *Journal of Finance*, **44** (1989), pp. 231–262.
9. Brennan, Michael J. "Taxes, Market Valuation, and Corporate Financial Policy," *National Tax Journal*, **25** (1970), pp. 417–427.
10. ——. "Capital Market Equilibrium with Divergent Borrowing and Lending Rates," *Journal of Financial and Quantitative Analysis*, **VI,** No. 5 (Dec. 1971), pp. 1197–1205.
11. Breeden, D., Gibbons M., and Litzenberger, R., "Empirical Tests of the Consumption-Oriented CAPM," *Journal of Finance*, **44,** No. 2 (1989), pp. 231–262.
12. Brenner, Menachem, and Subrahmanyam, Marti. "Intra-Equilibrium and Inter-Equilibrium Analysis in Capital Market Theory: A Clarification," *Journal of Finance*, **XXII,** No. 4 (Sept. 1977), pp. 1313–1319.
13. ——. "Portfolio Selection in an Economy with Marketability and Short Sales Restrictions," *Journal of Finance*, **XXXIII,** No. 2 (May 1978), pp. 589–601.
14. Chamberlain, G., and Rothschild, M. "Arbitrage, Factor Structure, and Mean–Variance Analysis on Large Asset Markets," *Econometrica*, **51** (1983), pp. 1281–1304.
15. Chen, N., Roll, R., and Ross, S. "Economic Forces and the Stock Market," *Journal of Business*, **59** (1986), pp. 386–403.
16. Cochrane, J. *Asset Pricing*. (Princeton University Press: Princeton NJ, 2001).
17. Connor, G. "A Unified Beta Pricing Theory," *Journal of Economic Theory*, **34** (1984), pp. 13–31.
18. Connor, G., and Korajczyk, R. "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, **15** (1986), pp. 373–394.
19. Constantinides, George M. "Admissible Uncertainty in the Intertemporal Asset Pricing Model," *Journal of Financial Economics*, **8,** No. 1 (March 1980), pp. 71–87.
20. Cornell, B. "The Consumption Based Asset Pricing Model: A Note on Potential Tests and Applications," *Journal of Financial Economics*, **9** (1981), pp. 103–108.
21. Dhrymes, Phoebus, Friend, Irwin, and Gultekin, Bulent. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory," *The Journal of Finance*, **39** (June 1984), pp. 323–346.
22. Dybvig, Philip H. "An Explicit Bound on Deviations from APT Pricing in a Finite Economy," *Journal of Financial Economics*, **12** (1983), pp. 483–496.
23. ——. "Distributional Analysis of Portfolio Choice," *The Journal of Business*, **61,** No. 2 (July 1988), pp. 369–393.
24. Dybvig, P., and Ross, S. "Yes, the APT Is Testable," *Journal of Finance*, **40** (1985), pp. 1173–1188.
25. Easley, David, and Jarrow, Robert A. "Consensus Beliefs Equilibrium and Market Efficiency," *The Journal of Finance*, **38,** No. 3 (June 1983), pp. 903–912.
26. Elton, Edwin J., and Gruber, Martin J. "The Multi-Period Consumption Investment Decision and Single Period Analysis," *Oxford Economic Papers*, **26** (Sept. 1974), pp. 180–195.
27. ——. *Finance as a Dynamic Process* (Englewood Cliffs, NJ: Prentice Hall, 1975).
28. ——. "Taxes and Portfolio Composition," *Journal of Financial Economics*, **6** (1978), pp. 399–410.
29. Errunza, Vihang, and Losq, Etienne. "International Asset Pricing under Mild Segmentation: Theory and Test," *The Journal of Finance*, **40,** No. 1 (March 1985), pp. 105–124.
30. Fama, Eugene. "Multi-Period Consumption-Investment Decision," *American Economic Review*, **60** (March 1970), pp. 163–174.
31. ——. "Risk, Return and Equilibrium," *Journal of Political Economy*, **79,** No. 1 (Jan.–Feb. 1971), pp. 30–55.
32. ——. "A Note on the Market Model and the Two-Parameter Model," *Journal of Finance*, **XXVIII,** No. 5 (Dec. 1973), pp. 1181–1185.
33. ——. *Foundations of Finance* (New York: Basic Books, 1976).

34. Fama, E., MacBeth, J., and Schwert, G. "Asset Returns and Inflation," *Journal of Financial Economics*, **5** (1977), pp. 115–146.
35. ——. "Inflation, Interest and Relative Prices," *Journal of Business*, **52** (1979), pp. 183–209.
36. Ferson, W. "Expected Real Interest Rates and Consumption in Efficient Financial Markets: Empirical Tests," *Journal of Financial and Quantitative Analysis*, **18** (1983), pp. 477–498.
37. Figlewski, Stephen. "Information Diversity and Market Behavior," *The Journal of Finance*, **37,** No. 1 (March 1982), pp. 87–102.
38. Foster, F. Douglas. "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution the Maximal R2," *The Journal of Finance*, **52,** No. 2 (June 1997), pp. 591–607.
39. Friend, Irwin, and Westerfield, Randolph. "Co-Skewness and Capital Assets Pricing," *The Journal of Finance*, **35,** No. 4 (Sept. 1980), pp. 897–914.
40. Friend, Irwin, Landskroner, Yoram, and Losq, Etienne. "The Demand for Risky Assets and Uncertain Inflation," *Journal of Finance*, **XXXI,** No. 5 (Dec. 1976), pp. 1287–1297.
41. Gibbons, M., and Ferson, W. "Testing Asset Pricing Models with Changing Expectations and an Unobservable Market Portfolio," *Journal of Financial Economics*, **14** (1985), pp. 217–236.
42. Gonedes, Nicholas. "Capital Market Equilibrium for a Class of Heterogeneous Expectations in a Two-Parameter World," *Journal of Finance*, **XXXI,** No. 1 (March 1976), pp. 1–15.
43. Grinblatt, Mark, and Titman, Sheridan. "Factor Pricing in a Finite Economy," *Journal of Financial Economics*, **12** (1983), pp. 497–507.
44. Grossman, S., and Shiller, R. "Consumption Correlatedness and Risk Measurement in Economies with Non-Traded Assets and Heterogeneous Information," *Journal of Financial Economics*, **10** (1982), pp. 195–210.
45. Grossman, S., Melino, A., and Shiller, R. "Estimating the Continuous-Time Consumption-Based Asset-Pricing Model," *Journal of Business and Economic Statistics*, **5** (1987), pp. 315–328.
46. Guiso, Luigi, Jappelli, Tullio, and Terlizzese, Daniele. "Income Risk, Borrowing Constraints, and Portfolio Choice," *The American Economic Review*, **86,** No. 1 (March 1996), pp. 158–172.
47. Hagerman, Robert, and Kim, Han. "Capital Asset Pricing with Price Level Changes," *Journal of Financial and Quantitative Analysis*, **XI,** No. 3 (Sept. 1976), pp. 381–391.
48. Hall, R. "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, **86** (1978), pp. 971–987.
49. Hansen, L., and Singleton, K. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, **50** (1982), pp. 1269–1286.
50. ——. "Stochastic Consumption, Risk Aversion, and the Temporary Behavior of Asset Returns," *Journal of Political Economy*, **91** (1983), pp. 249–265.
51. Hart, Oliver. "On the Existence of Equilibrium in a Securities Model," *Journal of Economic Theory*, **9,** No. 3 (Nov. 1974), pp. 293–311.
52. Heckerman, Donald. "Portfolio Selection and the Structure of Capital Asset Prices When Relative Prices of Consumption Goods May Change," *Journal of Finance*, **XXVII,** No. 1 (March 1972), pp. 47–60.
53. ——. "Reply to [52]," *Journal of Finance*, **XXVIII,** No. 5 (Dec. 1973), p. 1361.
54. Hilliard, Jimmy E. "Asset Pricing under a Subset of Linear Risk Tolerance Functions and Log-Normal Market Returns," *Journal of Financial and Quantitative Analysis*, **XV,** No. 5 (Dec. 1980), pp. 1041–1062.
55. Hogan, William, and Warren, James. "Toward the Development of an Equilibrium Capital-Market Model Based on Semi-Variance," *Journal of Financial and Quantitative Analysis*, **IX,** No. 1 (Jan. 1974), pp. 1–11.
56. Hopewell, Michael. "Comment on [88]: A Model of Capital Asset Risk," *Journal of Financial and Quantitative Analysis*, **VII,** No. 2 (March 1972), pp. 1673–1677.
57. Ibbotson, Roger, and Sinquefield, Rex. *Stocks, Bonds, Bills and Inflation: The Past and the Future* (Charlottesville, VA: Financial Analysts Research Foundation, 1982).
58. Ingersoll, Jonathan E., Jr. "Some Results in the Theory of Arbitrage Pricing," *Journal of Finance*, **39** (1984), pp. 1021–1039.

59. Jarrow, Robert. "Heterogeneous Expectations, Restrictions on Short Sales, and Equilibrium Asset Prices," *The Journal of Finance*, **35,** No. 5 (Dec. 1980), pp. 1105–1114.
60. Jobson, J., and Korkie, B. "Estimation for Markowitz Efficient Portfolios," *Journal of the American Statistical Association*, **75** (1980), pp. 544–554.
61. Jobson, J., and Korkie, R. "Potential Performance Tests of Portfolio Efficiency," *Journal of Financial Economics*, **10** (1982), pp. 433–466.
62. Kamoike, Osamu. "Portfolio Selection When Future Prices of Consumption Goods May Change: Comment on [36]," *Journal of Finance*, **XXVIII,** No. 5 (Dec. 1973), pp. 1357–1360.
63. Kandel, S. "On the Exclusion of Assets from Tests of the Mean–Variance Efficiency of the Market Portfolio," *Journal of Finance*, **39** (1984), pp. 63–75.
64. Kandel, S. "The Likelihood Ratio Test Statistic of Mean–Variance Efficiency without a Riskless Asset," *Journal of Financial Economics*, **13** (1984), pp. 575–592.
65. Kandel S., and Stambaugh, R. "On Correlations and the Sensitivity of Inferences about Mean–Variance Efficiency," *Journal of Financial Economics*, **18** (1987), pp. 61–90.
66. Keim, D. "Size Related Anomalies and Stock Return Seasonability: Further Empirical Evidence," *Journal of Financial Economics*, **12** (1983), pp. 13–32.
67. Korkie, Bob. "Comment: on [73]," *Journal of Financial and Quantitative Analysis*, **IX,** No. 5 (Nov. 1974), pp. 723–725.
68. Kraus, Alan, and Litzenberger, Robert. "Market Equilibrium in a Multi-Period State Preference Model with Logarithmic Utility," *Journal of Finance*, **XXX,** No. 5 (Dec. 1975), pp. 1213–1227.
69. ——. "Skewness Preference and the Valuation of Risk Assets," *Journal of Finance*, **XXXI,** No. 4 (Sept. 1976), pp. 1085–1100.
70. Kryzanowski, Lawrence, and Chau, To Hinh. "Asset Pricing Models When the Number of Securities Held Is Constrained: A Comparison and Reconciliation of the Mao and Levy Models," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1982), pp. 63–74.
71. Kumar, Prem. "Market Equilibrium and Corporation Finance: Some Issues," *Journal of Finance*, **XXIX,** No. 4 (Sept. 1974), pp. 1175–1188.
72. Landskroner, Yoram. "Nonmarketable Assets and the Determinants of the Market Price of Risk," *Review of Economics and Statistics*, **LIX,** No. 4 (Nov. 1977), pp. 482–514.
73. ——. "Intertemporal Determination of the Market Price of Risk," *Journal of Finance*, **XXXII,** No. 5 (Dec. 1977), pp. 1671–1681.
74. Lehari, David, and Levy, Haim. "The Capital Asset Pricing Model and the Investment Horizon," *Review of Economics and Statistics*, **LIX,** No. 1 (Feb. 1977), pp. 92–104.
75. Lettau, M., and Ludvigson, S. "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying," *Journal of Political Economy*, **109,** No. 6 (2001), pp. 1238–1287.
76. Levy, Haim. "The Capital Asset Pricing Model, Inflation, and the Investment Horizon: The Israeli Experience," *Journal of Financial and Quantitative Analysis*, **XV,** No. 3 (Sept. 1980), pp. 561–594.
77. Li, Q., Vassalou, M., and Xing, Y. "Sector Investment Growth Rates and the Cross Section of Equity Returns." *Journal of Business*, **89** (2006), pp. 1637–1665.
78. Lindenberg, Eric. "Imperfect Competition among Investors in Security Markets," Ph.D. dissertation, New York University, (1976).
79. ——. "Capital Market Equilibrium with Price Affecting Institutional Investors," in Edwin J. Elton and Martin J. Gruber (eds.), *Portfolio Theory 25 Years Later* (Amsterdam: North-Holland, 1979).
80. Lintner, John. "The Aggregation of Investors Diverse Judgments and Preferences in Purely Competitive Security Markets," *Journal of Financial and Quantitative Analysis*, **4,** No. 4 (Dec. 1969), pp. 347–400.
81. ——. "The Effect of Short Selling and Margin Requirements in Perfect Capital Markets," *Journal of Financial and Quantitative Analysis*, **VI,** No. 5 (Dec. 1971), pp. 1173–1195.
82. Litzenberger, R., and Ronn, E. "A Utility Based Model of Common Stock Returns," *Journal of Finance*, **41** (1986), pp. 67–92.

83. Long, John. "Stock Prices, Inflation, and the Term Structure of Interest Rates," *Journal of Financial Economics*, **1,** No. 2 (July 1974), pp. 131–170.
84. Losq, Etienne, and Chateau, John Peter D. "A Generalization of the CAPM Based on a Property of the Covariance Operator," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 5 (Dec. 1982), pp. 783–798.
85. Lucas, R. "Asset Prices in an Exchange Economy," *Econometrica*, **46** (1978), pp. 1429–1445.
86. Mayers, D. "Nonmarketable Assets and Capital Market Equilibrium under Uncertainty," in M. C. Jensen (ed.), *Studies in Theory of Capital Markets* (New York: Praeger, 1972).
87. Mayers, David. "Nonmarketable Assets and the Determination of Capital Asset Prices in the Absence of a Riskless Asset," *Journal of Business*, **46,** No. 2 (April 1973), pp. 258–267.
88. ——. "Nonmarketable Assets: Market Segmentation and the Level of Asset Prices," *Journal of Financial and Quantitative Analysis*, **XI,** No. 1 (March 1976), pp. 1–37.
89. Merton, Robert. "An Intertemporal Capital Asset Pricing Model," *Econometrica*, **41,** No. 5 (Sept. 1973), pp. 867–888.
90. Milne, Frank, and Smith, Clifford, Jr. "Capital Asset Pricing with Proportional Transaction Cost," *Journal of Financial and Quantitative Analysis*, **XV,** No. 2 (June 1980), pp. 253–266.
91. Ohlson, James. "Equilibrium in Stable Markets," *Journal of Political Economy*, **85,** No. 4 (Aug. 1977), pp. 859–864.
92. Paxson, Christina. "Borrowing Constraints and Portfolio Choice," *The Quarterly Journal of Economics*, **105,** No. 2 (May 1990), pp. 535–543.
93. Peles, Yoram. "A Note on Risk and the Theory of Asset Value," *Journal of Financial and Quantitative Analysis*, **VI,** No. 1 (Jan. 1971), pp. 643–647.
94. Pettit, R. Richardson, and Stanley, L. "Consumption-Investment Decisions with Transaction Costs and Taxes: A Study of the Clientele Effect of Dividends," *Journal of Financial Economics*, **5,** No. 3 (1979), pp. 551–572.
95. Pettit, R. Richardson, and Westerfield, Randolph. "A Model of Capital Asset Risk," *Journal of Financial and Quantitative Analysis*, **VII,** No. 2 (March 1972), pp. 1649–1668.
96. Rabinovitch, Ramon, and Owen, Joel. "Non-Homogeneous Expectations and Information in the Capital Asset Market," *Journal of Finance*, **XXXIII,** No. 2 (May 1978), pp. 575–587.
97. Reinganum, Marc R. "A New Empirical Perspective on the CAPM," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 4 (Nov. 1981), pp. 439–462.
98. Roberts, Gordon. "Endogenous Endowments and Capital Asset Prices," *Journal of Finance*, **XXX,** No. 1 (March 1975), pp. 155–162.
99. Roll, Richard, and Ross, Stephen. "An Empirical Investigation of Arbitrage Pricing Theory," *Journal of Finance* (Dec. 1980), pp. 1073–1105.
100. Rosenberg, B., and Guy, J. "Prediction of Beta from Investment Fundamentals," *Financial Analysts Journal*, **32** (1976), pp. 60–72.
101. Ross, Stephen. "Return, Risk, and Arbitrage," in I. Friend and J. Bickster (eds.), *Risks and Return in Finance* (Cambridge, MA: Ballinger, 1977).
102. ——. "The Capital Asset Pricing Model (CAPM), Short-Sale Restrictions and Related Issues," *Journal of Finance*, **XXXII,** No. 1 (March 1977), pp. 177–183.
103. ——. "Mutual Fund Separation in Financial Theory—The Separating Distributions," *Journal of Economic Theory*, **17,** No. 2 (April 1978), pp. 254–286.
104. ——. "The Current Status of the Capital Asset Pricing Model (CAPM)," *Journal of Finance*, **XXXIII,** No. 3 (June 1978), pp. 885–901.
105. Rubinstein, M. "The Valuation of Uncertain Income Streams and the Pricing of Options," *Bell Journal of Economics and Management Science*, **7** (1976), pp. 407–425.
106. Rubinstein, Mark. "The Strong Case for the Generalized Logarithmic Utility Model as the Premier Model of Financial Markets," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 551–571.
107. Samuelson, Paul. "Lifetime Portfolio Selection by Dynamic Stochastic Programming," *Review of Economics and Statistics*, **LI,** No. 3 (Aug. 1969), pp. 239–246.
108. Samuelson, Paul, and Merton, Robert. "Generalized Mean–Variance Tradeoffs for Best Perturbation Corrections to Approximate Portfolio Decisions," *Journal of Finance*, **XXIX,** No. 1 (March 1974), pp. 27–40.

109. Sandmo, Agnar. "Capital Risk, Consumption and Portfolio Choice," *Econometrica*, **37,** No. 4 (Oct. 1969), pp. 586–599.
110. Scholes, M., and Williams, J. "Estimating Betas from Nonsynchronous Data," *Journal of Financial Economics*, **5** (1977), pp. 309–327.
111. Shanken, J. "An Asymptotic Analysis of the Traditional Risk-Return Model," unpublished manuscript, School of Business Administration, University of California, Berkeley (1982).
112. ——. "Multi-Beta CAPM or Equilibrium-APT? A Reply," *Journal of Finance*, **40** (1985), pp. 1186–1189.
113. ——. "Multivariate Tests of the Zero-Beta CAPM," *Journal of Financial Economics*, **14** (Sept. 1985), pp. 327–348.
114. ——. "On Exclusion of Assets from Tests of the Mean–Variance Efficiency of the Market Portfolio: An Extension," *Journal of Finance*, **41** (1986), pp. 331–337.
115. ——. "A Posterior-Odds Ratio Approach to Testing Portfolio Efficiency," working paper, Graduate School of Management, University of Rochester (1986).
116. ——. "Testing Portfolio Efficiency When the Zero-Beta Rate Is Unknown: A Note," *Journal of Finance*, **41** (1986), pp. 269–276.
117. ——. "Multivariate Proxies and Asset Pricing Relations," *Journal of Financial Economics*, **18** (1987), pp. 91–110.
118. Shanken, Jay. "The Arbitrage Pricing Theory: Is It Testable?" *Journal of Finance*, **37** (1982), pp. 1129–1140.
119. Sharpe, William. *Portfolio Theory and Capital Markets* (New York: McGraw-Hill, 1970).
120. Siegel, Jeremy, and Warner, Jarold. "Indexation, the Risk-Free Asset, and Capital Market Equilibrium," *Journal of Finance*, **XXXII,** No. 4 (Sept. 1977), pp. 1101–1107.
121. Stapleton, Richard, and Subrahmanyam, Marti. "Multi-Period Equilibrium Asset Pricing Model," *Econometrica*, **46** (1977), pp. 1077–1096.
122. Stone, Bernell. "Systematic Interest-Rate Risk in a Two-Index Model of Returns," *Journal of Financial and Quantitative Analysis*, **IX,** No. 5 (Nov. 1974), pp. 709–721.
123. Viard, Alan D. "The Asset Pricing Effects of Fixed Holding Costs: An Upper Bound," *Journal of Financial and Quantitative Analysis*, **30,** No. 1 (March 1995), pp. 43–59.
124. Williams, Joseph. "Risk, Human Capital, and the Investor's Portfolio," *Journal of Business*, **51,** No. 1 (Jan. 1978), pp. 65–89.

15

Empirical Tests of Equilibrium Models

In the two previous chapters we stressed the fact that the construction of a theory necessitates a simplification of the phenomena under study. To understand and model any process, elements in the real world are simplified or assumed away. While a model based on simple assumptions can always be called into question because of these assumptions, the relevant test of how much damage has been done by the simplification is to examine the relationship between the predictions of the model and observed real-world phenomena. In our case, the relevant test is how well the simple capital asset pricing model (CAPM), or perhaps some other general equilibrium model, describes the behavior of actual capital markets.

The principle is easily stated and intuitively appealing. However, it opens up a new series of problems. Namely, how does one design meaningful empirical tests of a theory? In particular, how can one test the CAPM or any of its numerous variants? In this chapter we review several of the tests of the general equilibrium models that have been presented in the literature. In doing so, we discuss many of the problems encountered in designing these tests. Finally, we discuss fundamental work by Roll (1977) that suggests certain problems with all of the tests of general equilibrium models and opens up the area to further questions.

THE MODELS—EX ANTE EXPECTATIONS AND EX POST TESTS

Most tests of general equilibrium models deal with either the standard CAPM or the zero-beta (two-factor) form of a general equilibrium model. The basic CAPM can be written as

$$E(R_i) = R_F + \beta_i\left[E(R_M) - R_F\right]$$

The no lending or borrowing version, often called the two-factor model, can be written as

$$E(R_i) = E(R_Z) + \beta_i\left[E(R_M) - E(R_Z)\right]$$

Recall that $E(R_Z)$ is the expected return on the minimum-variance portfolio that is uncorrelated with the market portfolio.

Notice that these models are formulated in terms of expectations. All variables are expressed in terms of future values. The relevant beta is the future beta on the security. Furthermore, both the return on the market and the return on the minimum-variance zero-beta portfolio are expected future returns.

Because large-scale systematic data on expectations do not exist, almost all tests of the CAPM have been performed using ex post or observed values for the variables. This raises the logical question of how one justifies testing an expectational model in terms of realizations.

There are two lines of defense that have commonly been used by researchers. The simpler defense is to argue that expectations are on average and, on the whole, correct. Therefore, over long periods of time, actual events can be taken as proxies for expectations.

The more complex defense starts by assuming that security returns are linearly related to the return on a market portfolio (a version of the single-index model of Chapter 7). This model, called the market model, can be written as

$$\tilde{R}_{it} = \alpha_i + \beta_i \tilde{R}_{Mt} + \tilde{e}_{it} \tag{15.1}$$

The squiggle over a variable indicates that the variable is random.

The expected value of the return on security *i* is

$$E(R_i) = \alpha_i + \beta_i E(R_M)$$

Thus

$$E(R_i) - \alpha_i - \beta_i E(R_M) = 0$$

Adding this equation to the right-hand side of Equation (15.1) and rearranging yields

$$\tilde{R}_{it} = E(R_i) + \beta_i \left[\tilde{R}_{Mt} - E(R_M)\right] + \tilde{e}_{it}$$

The simple form of the CAPM model is

$$E(R_i) = R_F + \beta_i \left[E(R_M) - R_F\right]$$

Substituting the expression for $E(R_i)$ into the previous equation and simplifying,

$$\tilde{R}_{it} = R_F + \beta_i \left(\tilde{R}_{Mt} - R_F\right) + \tilde{e}_{it} \tag{15.2}$$

Testing a model of this form with ex post data seems appropriate. However, notice that there are three assumptions behind this model:

1. The market model holds in every period.

2. The CAPM model holds in every period.

3. The beta is stable over time.

A test of this model on ex post data is really a simultaneous test of all three of these hypotheses.

The reader should note that if one had used the two-factor model instead of the Sharpe–Lintner–Mossin form, we would have found

$$\tilde{R}_{it} = \tilde{R}_{Zt} + \beta_i \left(\tilde{R}_{Mt} - \tilde{R}_{Zt}\right) + \tilde{e}_{it} \tag{15.3}$$

rather than Equation (15.2). As in the previous case, a test of this model is really a simultaneous test of three hypotheses: the zero-beta version of the CAPM model holds in every period, the market model holds in every period, and beta is stable over time. However, making these assumptions does express the model in terms of realized returns.

EMPIRICAL TESTS OF THE CAPM

There has been a huge amount of empirical testing of the standard form and the two-factor form of the CAPM model. A discussion of all empirical work would require a volume by

itself. The approach we have adopted is to review the hypotheses that should be tested, to review some of the early work on testing the CAPM, then to discuss briefly a few of the problems inherent in any test of the CAPM. Finally, we review, in more detail, some of the more rigorous tests.

Some Hypotheses of the CAPM

Certain hypotheses can be formulated that should hold whether one believes in the simple CAPM or the two-factor general equilibrium model.

- The first is that higher risk (beta) should be associated with a higher level of return.
- The second is that return is linearly related to beta; that is, for every unit increase in beta, there is the same increase in return.
- The third is that there should be no added return for bearing nonmarket risk.

In addition, if some form of general equilibrium model holds, then investing should constitute a fair game with respect to it. That is, deviations of a security or portfolio from equilibrium should be purely random, and there should be no way to use these deviations to earn an excess profit.

In addition to the hypotheses common to both the standard and the two-factor form of the CAPM, we can formulate hypotheses that attempt to differentiate between these general equilibrium models. In particular, the standard version implies that the security market line, drawn in return beta space, should have an intercept of R_F and a slope of $(\bar{R}_M - R_F)$, while the two-factor version requires that it should have an intercept of $\bar{R}_Z$ and a slope of $(\bar{R}_M - \bar{R}_Z)$.

A Simple Test of the CAPM

Before we become involved in a discussion of the history and methodology of tests of the CAPM model, it seems worthwhile examining the results of a simple test of the CAPM to see if, over long periods of time, higher return has been associated with higher risk (as measured by beta). Sharpe and Cooper (1972) examined whether following alternative strategies, with respect to risk over long periods of time, would produce returns consistent with modern capital theory. To get portfolios with different betas, they divided stocks into deciles once a year on the basis of the beta of each security.[1] To be more precise, beta at a point in time was measured using 60 months of previous data. Once a year, for each year 1931–1967, all New York Stock Exchange stocks were divided into deciles based on their rank by beta. An equally weighted portfolio was formed of the stocks that comprised each decile. A strategy consisted of holding the stocks of a particular decile over the entire period. The stocks one holds change both because of the reinvestment of dividends and because the stocks that make up a particular decile change as the decile's composition is revised once a year. Notice that the strategy outlined by Sharpe and Cooper could actually be followed by an investor. Each year the investor divides stocks into deciles by beta based on the previous five years' (60 months) returns. If investors want to pursue the high-beta strategy, they simply divide their funds equally among the stocks in the highest beta decile. They do this every year and observe the outcomes. Table 15.1 shows what would have happened, on average, if an investor had done this each year from 1931 to 1967.

Although the relationship between strategy and return is not perfect, it is very close. In general, stocks with higher betas have produced higher future returns. In fact, the rank

[1]The measure of beta they used was analogous to the standard beta computed by regressing the returns from any security against the market. The difference was that dividends were excluded both from the market and the stocks' return. The authors found the coefficient of determination between standard beta and their measure was 0.996.

Table 15.1 Average Returns and Betas on Portfolios Ranked by Betas

Strategy	Average Return	Portfolio Beta
10	22.67	1.42
9	20.45	1.18
8	19.116	1.14
7	21.77	1.24
6	18.49	1.06
5	19.13	0.98
4	18.88	1.00
3	14.99	0.76
2	14.63	0.65
1	11.58	0.58

correlation coefficient between strategy and return is over 0.93, which is statistically significant at the 0.01 level. Similarly, buying stocks with higher forecast beta would lead to holding portfolios with higher realized betas. The rank correlation between strategy and beta is 95%, which is significant at the 0.01 level.

The next logical step is to examine the relationship between the return that would have been earned and the risk (beta) from following alternative strategies. Figure 15.1 from Sharpe and Cooper (1972) shows this relationship. The equation of this graph is

$$\overline{R}_i = 5.54 + 12.75\beta_i$$

More than 95% of the variation in expected return is explained by differences in beta. Thus beta has explained a very significant portion of the difference in return between these portfolios (strategies).

Sharpe and Cooper's work presents rather clear and easily interpreted evidence that, as general equilibrium theory suggests, there is a positive relationship between return and beta. Furthermore, an examination of Figure 15.1 provides confidence that the relationship is both strong and linear. The intercept of 5.54 is considerably higher than the riskless rate (rate on Treasury bills), which was below 2% during this period. This lends

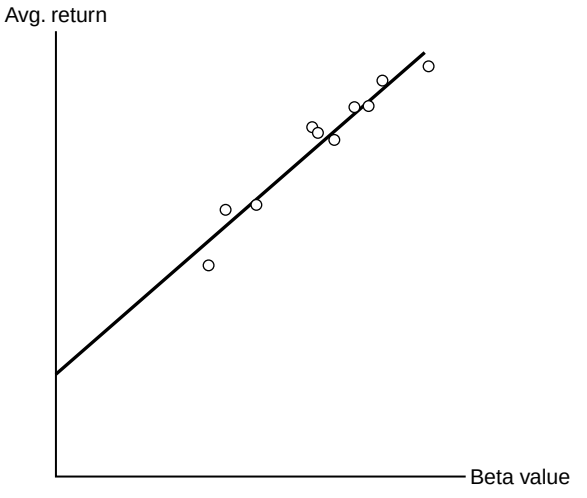


Figure 15.1 Estimated security market line.

support to the two-factor form of the CAPM. Let us now turn to some more sophisticated tests of the CAPM.

Some Early Empirical Tests

Most of the early empirical tests of the CAPM involved the use of a time series (first pass) regression to estimate betas and the use of a cross-sectional (second pass) regression to test the hypotheses we derived from the CAPM model. To make this more concrete, let us turn to an early empirical study of the CAPM performed by Lintner and reproduced in Douglas (1968). Lintner first estimated beta for each of the 301 common stocks in his sample. He estimated beta by regressing each stock's yearly return against the average return for all stocks in the sample using data from 1954 to 1963. The first-pass regression had the form

$$R_{it} = \alpha_i + b_i R_{Mt} + e_{it}$$

where b_i (the regression coefficient) was the estimate of the true beta for stock *i*. Lintner then performed the second-pass cross-sectional regression

$$\bar{R}_i = a_1 + a_2 b_i + a_3 S_{ei}^2 + \eta_i$$

where S_{ei}^2 is the residual variance from the first-pass regression (the variance of e_i). Each parameter of this model has a theoretical value: a_3 should be equal to zero, a_1 should be equal to either R_F or $\bar{R}_Z$, and a_2 should be equal to either $\bar{R}_M - R_F$ or $\bar{R}_M - \bar{R}_Z$, according to the form of the CAPM that is being tested.[2] The values he obtained were

$$a_1 = 0.108$$
$$a_2 = 0.063$$
$$a_3 = 0.237$$

These results seem to violate the CAPM.[3] The term representing residual risk was statistically significant and positive. The intercept term a_1 would seem to be larger than any reasonable estimate of either R_F or $\bar{R}_Z$, and a_2, although statistically significant, has a value slightly lower than we could reasonably expect. Douglas (1968) employed a similar methodology and found results that were similar to Lintner's.

Tests of Black, Jensen, and Scholes

Miller and Scholes (1972) in a classic article show that the anomalous results reported by Lintner may be an artifact of a number of statistical issues, most notably that the beta measured in the first-pass regression is only an estimate of the true beta. Using the estimate of beta in the second-pass cross-sectional regression will lead to an errors-in-variables problem implying a bias correlated with the standard error of the estimate of beta, sufficient to explain the positive coefficient on residual variance in the second-pass regression Lintner reports. To mitigate this errors-in-variables problem, Black, Jensen, and Scholes (1972) first form portfolios based on prior estimates of beta. The asset pricing model holds for portfolios as well as for individual securities, and so in the first stage, betas are estimated for each of 10 portfolios.

---

[2]These theoretical values arise from Equations (15.2) and (15.3).

[3]Both a_2 and a_3 are statistically different from zero at the 0.01 level. The *t* values for these coefficients are 6.9 and 6.8, respectively.

Table 15.2 Tests of the CAPM as Reported by Black, Jensen, and Scholes (1972)

	β	Excess Return[a]	α_i Intercept	ρ^b
1	1.561	0.0213	−0.0829	0.963
2	1.384	0.0177	−0.1938	0.988
3	1.248	0.0171	−0.0649	0.988
4	1.163	0.0163	−0.0167	0.991
5	1.057	0.0145	−0.0543	0.992
6	0.923	0.0137	0.0593	0.983
7	0.853	0.0126	0.0462	0.985
8	0.753	0.0115	0.0812	0.979
9	0.629	0.0109	0.1968	0.956
10	0.490	0.0091	0.2012	0.898
Market	1.000	0.0142		

[a]On monthly terms, 0.0213 should be read as 2.13% return per month. Excess return is average return on the portfolio minus the risk-free rate.

[b]Correlation coefficient.

The results are shown in Table 15.2. If the zero-beta model rather than the standard model holds, then the intercept is the difference between $\bar{R}_Z$ and $\bar{R}_F$ times one minus beta, or

$$\alpha_i = \left(\bar{R}_Z - R_F\right)\left(1 - \beta_i\right)$$

As shown in Chapter 14, $\bar{R}_Z$ should be larger than R_F. Thus ($\bar{R}_Z - R_F$) should be positive. Therefore, if β_i is less than 1, α_i should be positive, and if β_i is greater than 1, α_i should be negative. This is exactly what the empirical results show. Black, Jensen, and Scholes repeat these tests for four subperiods and find, by and large, the same type of behavior we have described for the overall period.

In a second-pass regression, average returns in excess of Treasury bill returns are regressed on these estimates of beta to find

$$\bar{R}_i - R_F, = 0.00359 + 0.01080\ \beta_i, \qquad \rho^2 = 0.98$$

The positive value of the intercept that emerges is evidence in support of the two-factor model. The high percentage of the variation in returns explained (98%) seems to show that a straight line describes returns very well, as predicted by the theory. A recent paper by Lewellen, Nagel, and Shanken (2010) argues that the high explanatory power of the second-pass cross-sectional regressions may be an artifact of constructing portfolios on the basis of the same risk factor used in the first-pass regression. They show that this apparent explanatory power is reduced once industry representation rather than risk factors is used to form portfolios. Fama and French (1992) come to similar conclusions using portfolios organized by size, book to market, as well as beta and conclude that the relation between beta and average return is flat, even when beta is the only explanatory variable. Roll and Ross (1994) argue that this is an artifact of using ordinary least squares in the cross-sectional second-pass regression. They argue that the relationship between average returns and beta is retrieved once heteroskedasticity and cross-sectional residual correlation is accounted for using generalized least squares instead of the more usual ordinary least squares in the second-pass cross-sectional regression.[4]

[4]Nevertheless, Lewellen, Nagel, and Shanken (2010) argue that a high percentage of variance explained (in a second-pass generalized least squares context) is the appropriate test of the theory, because this percentage is proportional to the square of the implied Sharpe ratio of the optimal portfolio of assets and is thus related to the T^2 measure discussed later.

The approach of using portfolios in place of individual securities to alleviate the errors-in-variables problem has become standard in the empirical asset pricing literature. It is an important element in the Fama and MacBeth (1973) methodology to test the CAPM.

Tests of Fama and McBeth

Fama and McBeth formed 20 portfolios of securities to estimate betas from a first-pass regression. They then performed one second-pass cross-sectional regression for each month subsequent to the estimation period over the time period 1935–1968. The equation they tested was

$$\tilde{R}_{it} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\beta_i - \hat{\gamma}_{2t}\beta_i^2 + \hat{\gamma}_{3t}S_{ei} + \eta_{it} \tag{15.4}$$

By estimating this equation (in cross section) for each month, it is possible to study how the parameters change over time.

This form of the equation allows the test of a series of hypotheses regarding the CAPM. The tests are as follows:

1. $E(\hat{\gamma}_{3t}) = 0$, or residual risk does not affect return.
2. $E(\hat{\gamma}_{2t}) = 0$, or there are no nonlinearities in the security market line.
3. $E(\hat{\gamma}_{1t}) > 0$, that is, there is a positive price of risk in the capital markets.

If both $E(\hat{\gamma}_{2t})$ and $E(\hat{\gamma}_{3t})$ are not different from zero, we can also examine both $E(\hat{\gamma}_{0t})$ and $E(\hat{\gamma}_{1t})$ to see whether the standard CAPM or zero-beta model is a better description of market returns.

Finally, we can examine all of the coefficients and the residual term to see if the market operates as a fair game. If the market is a fair game, then there is no way that one should be able to use knowledge about the value of the parameters in previous periods to make an excess return. For example, if the standard CAPM or the zero-beta model holds, then, regardless of the prior values of γ_{2t} and γ_{3t}, each of their expected values at time $t + 1$ should be zero. Furthermore, if the zero-beta model is the best description of general equilibrium, then deviations of $\hat{\gamma}_{0t}$ from its mean $E(R_Z)$ and $\hat{\gamma}_{1t}$ from its mean $E(R_M) - E(R_Z)$ are random, regardless of what happened at time period $t - 1$ or any earlier time period. If the simple form of the CAPM holds the same, statements should be true with R_F substituted for $\bar{R}_Z$.

Fama and MacBeth have estimates of $\hat{\gamma}_{0t}$, $\hat{\gamma}_{1t}$, $\hat{\gamma}_{2t}$, and $\hat{\gamma}_{3t}$ and η_{it} for each month over the period January 1935–June 1968. The average value of any $\hat{\gamma}_{it}$ (denoted by $\bar{\hat{\gamma}}_i$) can be found simply by averaging the individual values, and this mean can be tested to see if it is different from zero.[5]

Table 15.3 from Fama and MacBeth (1973) presents the results of estimating Equation (15.4) and several variations of it over the full time period of 1935–1968, as well as for several subperiods. Notice that they have estimated the full Equation (15.4), as well as forms of the equation with all values of $\hat{\gamma}_{2t}$ and $\hat{\gamma}_{3t}$ both separately and simultaneously forced to zero. If both theory and empirical evidence indicate that one or more variables have no influence on an equation, better estimates of the remaining coefficients can be made when these influences do not enter the estimating equation. For example, theory and

[5]The statistical significance of each parameter can be found by calculating the standard deviation of the mean and testing to see if the mean is a significant number of standard deviations from zero. From the central limit theorem, the mean is normally distributed, with standard deviation equal to the standard deviation of the $\hat{\gamma}_{it}$s divided by the square root of the number of observations on $\hat{\gamma}_{it}$.

Table 15.3 Tests of the Two-Parameter Model

	Statistic																	
Period	$\bar{\hat{\gamma}}_0$	$\bar{\hat{\gamma}}_1$	$\bar{\hat{\gamma}}_2$	$\bar{\hat{\gamma}}_3$	$\overline{\hat{\gamma}_0 - R_F}$	$s(\hat{\gamma}_0)$	$s(\hat{\gamma}_1)$	$s(\hat{\gamma}_2)$	$s(\hat{\gamma}_3)$	$\hat{\rho}_0(\hat{\gamma}_0 - R_F)$	$\hat{\rho}(\hat{\gamma}_1)$	$\hat{\rho}_0(\hat{\gamma}_2)$	$\hat{\rho}_0(\hat{\gamma}_3)$	$t(\bar{\hat{\gamma}}_0)$	$t(\bar{\hat{\gamma}}_1)$	$t(\bar{\hat{\gamma}}_2)$	$t(\bar{\hat{\gamma}}_3)$	$t(\hat{\gamma}_0 - R_F)$	$\hat{\rho}^2$	$s(\rho^2)$
	$R_{it} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\beta_i + \eta_{it}$																			
Panel A																				
1935–6/1968	0.0061	0.0085			0.0048	0.038	0.066			0.15	0.02			3.24	2.57			2.55	0.29	0.30
1935–1945	0.0039	0.0163			0.0037	0.052	0.098			0.10	−0.03			0.86	1.92			0.82	0.29	0.29
1946–1955	0.0087	0.0027			0.0078	0.026	0.041			0.18	0.07			3.71	0.70			3.31	0.31	0.32
1956–6/1968	0.0060	0.0062			0.0034	0.030	0.044			0.27	0.15			2.45	1.73			1.39	0.28	0.29
1935–1940	0.0024	0.0109			0.0023	0.064	0.116			0.07	−0.09			0.32	0.79			0.31	0.23	0.30
1941–1945	0.0056	0.0229			0.0054	0.034	0.069			0.23	0.15			1.27	2.55			1.22	0.37	0.28
1946–1950	0.0050	0.0029			0.0044	0.031	0.047			0.20	0.04			1.27	0.48			1.10	0.39	0.33
1951–1955	0.0123	0.0024			0.0111	0.019	0.035			0.20	0.08			5.06	0.53			4.56	0.24	0.29
1956–1960	0.0148	−0.0059			0.0128	0.020	0.034			0.37	0.18			5.68	−1.37			4.89	0.22	0.31
1961–6/1968	0.0001	0.0143			−0.0029	0.034	0.048			0.22	0.09			0.03	2.81			−0.80	0.32	0.27
	$R_{it} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\beta_i + \hat{\gamma}_{2t}\beta_i^2 + \eta_{it}$																			
Panel B																				
1935–6/1968	0.0049	0.0105	−0.0008		0.0036	0.052	0.118	0.056		0.03	−0.11	−0.11		1.92	1.79	−0.29		1.42	0.32	0.31
1935–1945	0.0074	0.0079	0.0040		0.0073	0.061	0.139	0.074		−0.10	−0.31	−0.21		1.39	0.65	0.61		1.36	0.32	0.30
1946–1955	−0.0002	0.0217	−0.0087		−0.0012	0.036	0.095	0.034		0.04	0.00	0.00		−0.07	2.51	−2.83		−0.38	0.36	0.32
1956–6/1968	0.0069	0.0040	0.0013		0.0043	0.054	0.116	0.053		0.17	0.07	0.03		1.56	0.42	0.29		0.97	0.30	0.30
1935–1940	0.0013	0.0141	−0.0017		0.0012	0.069	0.160	0.075		−0.13	−0.36	−0.35		0.16	0.75	−0.19		0.14	0.24	0.30
1941–1945	0.0148	0.0004	0.0108		0.0146	0.050	0.111	0.073		−0.04	−0.19	−0.04		2.28	0.03	1.15		2.24	0.39	0.29
1946–1950	−0.0008	0.0152	−0.0051		−0.0015	0.037	0.104	0.032		0.14	0.04	0.00		−0.18	1.14	−1.24		−0.32	0.44	0.32
1951–1955	0.0004	0.0281	−0.0122		−0.0008	0.030	0.085	0.035		−0.17	−0.14	−0.01		0.10	2.55	−2.72		−0.20	0.28	0.29
1956–1960	0.0128	−0.0015	−0.0020		0.0108	0.030	0.072	0.029		0.35	0.11	0.26		3.38	−0.16	−0.54		2.84	0.25	0.31
1961–6/1968	0.0029	0.0077	0.0034		−0.0000	0.066	0.138	0.064		0.14	0.06	−0.01		0.42	0.53	0.51		−0.01	0.34	0.29

(*Continued*)

Table 15.3 (*Continued*)

Period	Statistic																		
	$\bar{\gamma}_0$	$\bar{\gamma}_1$	$\bar{\gamma}_2$	$\bar{\gamma}_3$	$\overline{\bar{\gamma}_0 - R_F}$	$s(\hat{\gamma}_0)$	$s(\hat{\gamma}_1)$	$s(\hat{\gamma}_2)$	$s(\hat{\gamma}_3)$	$\hat{\rho}_0(\hat{\gamma}_0 - R_F)$	$\hat{\rho}(\hat{\gamma}_1)$	$\hat{\rho}_0(\hat{\gamma}_2)$	$\hat{\rho}_0(\hat{\gamma}_3)$	$t(\bar{\gamma}_0)$	$t(\bar{\gamma}_1)$	$t(\bar{\gamma}_2)$	$t(\bar{\gamma}_3)$	$t(\bar{\gamma}_0 - R_F)$	$\hat{\rho}^2$	$s(\rho^2)$
$R_{it} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\beta_i + \hat{\gamma}_{3t}s_{ei} + \eta_{it}$																			
Panel C																			
1935–6/1968	0.0054	0.0072		0.0198	0.0041	0.052	0.065		0.868	0.04	−0.12		−0.04	2.10	2.20		0.46	1.59	0.32	0.31
1935–1945	0.0017	0.0104		0.0841	0.0015	0.073	0.083		0.921	−0.00	−0.26		−0.08	0.26	1.41		1.05	0.24	0.32	0.31
1946–1955	0.0110	0.0075		−0.1052	0.0100	0.032	0.056		0.609	0.08	0.02		−0.20	3.78	1.47		−1.89	3.46	0.34	0.32
1956–6/1968	0.0042	0.0041		0.0633	0.0016	0.040	0.052		0.984	0.12	0.08		0.03	1.28	0.96		0.79	0.50	0.30	0.29
1935–1940	0.0036	0.0119		−0.0170	0.0035	0.082	0.105		0.744	−0.03	−0.26		−0.18	0.37	0.97		−0.19	0.36	0.25	0.30
1941–1945	−0.0006	0.0085		0.2053	−0.0009	0.061	0.052		1.091	0.07	−0.29		−0.02	−0.08	1.25		1.46	−0.11	0.41	0.30
1946–1950	0.0069	0.0081		−0.0920	0.0062	0.034	0.066		0.504	0.14	0.06		−0.02	1.56	0.95		−1.41	1.40	0.42	0.33
1951–1955	0.0150	0.0069		−0.1185	0.0138	0.029	0.043		0.702	0.06	−0.18		−0.32	4.05	1.24		−1.31	3.72	0.27	0.29
1956–1960	0.0127	−0.0081		−0.0728	0.0107	0.037	0.045		1.164	0.15	0.15		0.21	2.68	−1.40		0.48	2.26	0.26	0.30
1961–6/1968	−0.0014	0.0122		0.0570	−0.0044	0.042	0.055		0.850	0.10	0.00		−0.19	−0.32	2.12		0.64	−0.98	0.33	0.27
$R_{it} = \hat{\gamma}_{0t} + \hat{\gamma}_{1t}\beta_i + \hat{\gamma}_{2t}\beta_i^2 + \hat{\gamma}_{3t}s_{ei} + \eta_{it}$																			
Panel D																			
1935–6/1968	0.0020	0.0114	−0.0026	0.0516	0.0008	0.075	0.123	0.060	0.929	−0.09	−0.09	−0.12	−0.10	0.55	1.85	−0.86	1.11	0.20	0.34	0.31
1935–1945	0.0011	0.0118	−0.0009	0.0817	0.0010	0.103	0.146	0.079	1.003	−0.20	−0.23	−0.24	−0.15	0.13	0.94	−0.14	0.94	0.11	0.34	0.31
1946–1955	0.0017	0.0209	−0.0076	−0.0378	0.0008	0.042	0.096	0.038	0.619	−0.10	−0.00	−0.01	−0.20	0.44	2.39	−2.16	−0.67	0.20	0.36	0.32
1956–6/1968	0.0031	0.0034	−0.0000	0.0966	0.0005	0.065	0.122	0.055	1.061	0.12	0.03	0.01	−0.05	0.59	0.34	−0.00	1.11	0.10	0.32	0.29
1934–1940	0.0009	0.0156	−0.0029	0.0025	0.0008	0.112	0.171	0.085	0.826	−0.16	−0.23	−0.26	−0.12	0.07	0.78	−0.29	0.03	0.06	0.26	0.30
1941–1945	0.0015	0.0073	0.0014	0.1767	0.0012	0.092	0.109	0.072	1.181	−0.28	−0.21	−0.22	−0.18	0.12	0.52	0.15	1.16	0.10	0.43	0.31
1946–1950	0.0011	0.0141	−0.0040	−0.0313	0.0004	0.047	0.106	0.042	0.590	−0.10	0.03	−0.01	−0.12	0.18	1.03	−0.73	−0.41	0.07	0.44	0.33
1951–1955	0.0023	0.0277	−0.0112	−0.0443	0.0011	0.037	0.085	0.034	0.651	−0.11	−0.13	−0.01	−0.28	0.48	2.53	−2.54	−0.53	0.23	0.29	0.30
1956–1960	0.0103	−0.0047	−0.0020	0.0979	0.0083	0.049	0.078	0.032	1.286	−0.16	0.19	−0.01	0.02	1.63	−0.47	−0.49	0.59	1.31	0.28	0.30
1961–6/1968	−0.0017	0.0088	0.0013	0.0957	−0.0046	0.073	0.144	0.066	0.887	0.20	0.00	0.01	−0.15	−0.21	0.58	0.19	1.02	−0.60	0.35	0.29

Source: Eugene F. Fama and James D. MacBeth (1974).

the initial empirical results (as we will see) indicate that neither β^2 nor residual risk affect return. Therefore better estimates of the effect of beta on return can be made when these variables are excluded because the coefficient on beta will not be affected by the multicollinearity between beta and beta square and between beta and residual risk.

Examining panels *C* and *D* of Table 15.3 reveals that, when measured over the entire period, $\bar{\hat{\gamma}}_3$ is small and is not statistically different from zero. Furthermore, when we examine it over several subperiods, we find that it remains small in each subperiod, is not significantly different from zero, and, in fact, exhibits different signs in different subperiods. We can safely conclude that residual risk has no effect on the expected return of a security. However, it is still possible that the market does not constitute a fair game with respect to any information contained in $\hat{\gamma}_{3t}$. That is, it is possible that the fact that $\hat{\gamma}_{3t}$ differs from zero in any period gives us insight into what its value (and, therefore, returns) will be next period. The easiest way to test this is to examine the correlation of $\hat{\gamma}_{3t}$ in one period with its value in the prior period, where the mean of all periods is assumed to be zero. Panels *C* and *D* show that the value of this correlation coefficient $[\rho_0(\gamma_3)]$ is close to zero and not statistically significant.[6] Fama and MacBeth also compute the correlation between $\hat{\gamma}_{3t}$ and its prior value for lags of more than one period. They find, once again, that there is no usable information contained in $\hat{\gamma}_{3t}$.

The results of Fama and MacBeth are opposite to those of Lintner and Douglas regarding the importance of residual risk. The earlier discussion provides a clue. Recall that Miller and Scholes showed that if beta had a large sampling error, then residual risk served as a proxy for true beta. Fama and MacBeth have much less sampling error than Lintner and Douglas because of their use of portfolios. When beta is estimated more accurately, residual risk no longer shows up as important.

The results, with respect to $\hat{\gamma}_{2t}$, are very similar. Examining panels *B* and *D*, we see that $\bar{\hat{\gamma}}_2$ is small, is not statistically significant, and changes sign over alternative subperiods. Furthermore, an examination of the correlation of $\hat{\gamma}_{2t}$ with its previous value (with means assumed to be zero) shows that there is no information contained in individual values of $\hat{\gamma}_{2t}$. Thus the beta squared term does not affect the expected return on securities, nor does its coefficient contain information with respect to an investment strategy.

Having concluded that neither beta squared nor residual risk has an influence on returns, the correct form of the equation to examine for further tests is that displayed in panel *A*.

Fama and MacBeth examine the performance of $\bar{\hat{\gamma}}_1$ for the entire period and conclude that there is evidence that the relationship between expected return and beta is positive as well as linear. Furthermore, by testing the correlation of the difference between $\hat{\gamma}_{1t}$ and its mean with prior values of the same variable, they show that difference in $\hat{\gamma}_{1t}$ from its mean cannot be employed to produce a better forecast of a future value of $\hat{\gamma}_{1t}$ than simply using the mean.

Fama and MacBeth find that $\bar{\hat{\gamma}}_0$ is generally greater than R_F, and over the entire period, $\bar{\hat{\gamma}}_1$ is statistically significantly greater than zero. In addition, they find that $\bar{\hat{\gamma}}_1$ is generally less than $\bar{R}_M - R_F$. The fact that $\bar{\hat{\gamma}}_0$ is substantially greater than R_F and that $\bar{\hat{\gamma}}_1$ is substantially less than $\bar{R}_M - R_F$ would seem to indicate that the zero-beta model is more consistent with equilibrium conditions than is the simple CAPM.[7]

---

[6]Fama and MacBeth point out that the standard deviation of the correlation coefficient can be approximated by 1 divided by the square root of the number of observations, or 0.05 for the overall period, 0.09 for the 10-year subperiod, and 0.13 for the 5-year subperiod.

[7]A warning is in order. Roll (1985) demonstrates that this difference could be due to the choice of a market index, and Fama (1976) also indicates that this might be true.

Before finishing our discussion of these tests, one more point is worth mentioning. If the equilibrium model describes market conditions, then an individual security's deviation from the model should contain no information. That is, a positive residual value for any one stock at any moment in time should convey no information about the differential performance of that stock (from the expected value produced by the model) in future periods. For this to be true, there should be no correlation (with any lag) between the residuals in Equation (15.4). This is, in fact, what Fama and MacBeth found.

Extensions of Fama and MacBeth

It is fair to say that the Fama and MacBeth paper is one of the most influential papers written on the empirical implications of the asset pricing model. Virtually every subsequent paper uses one or more of the elements it introduces. Key to all subsequent empirical studies of the asset pricing model is the idea that time series data are used to identify risk exposure (beta or other measures of risk), while cross-sectional differences identify (possibly time variant) risk premia. The result that a residual value for any one stock at any moment should contain no information about the differential performance in future periods is a key insight that has led to advanced estimation procedures robust to the kind of nonnormality issues raised by Miller and Scholes. One tempting implication of Fama and MacBeth's findings is that because mean–variance efficiency of the market implies the CAPM, validation of the linear equilibrium pricing model is an indirect test of mean–variance efficiency. We shall discuss Roll's (1977) critique of this interpretation, which has prompted a new concern for understanding the power of these tests to identify model failures when they occur.

While the Fama and MacBeth approach makes intuitive sense and is easy to implement, econometricians came to believe that there may be more powerful tests of the model. A few years after the publication of the original Fama and MacBeth paper, Gibbons (1982) employs the fact that the CAPM places a nonlinear restriction on a set of N regression equations, one for each security. More specifically, we know that the market model requires

$$R_{it} = \alpha_i + \beta_i\left(R_{mt}\right) + e_{it} \tag{15.5}$$

If the market model and the CAPM hold simultaneously, then

$$R_{it} = \gamma_1\left(1-\beta_i\right) + \beta_i\left(R_{mt}\right) + e_{it}$$

or

$$\alpha_i = \gamma_1\left(1-\beta_i\right) \tag{15.6}$$

where γ_1 is a constant for all securities but may vary from period to period. For the standard form of the CAPM, γ_1 should equal R_F, and for the zero-beta form, γ_1 should equal $\bar{R}_Z$, which should be larger than R_F. Now a set of N equations (one for each security) like (15.5) can be estimated simultaneously.

The same set of equations can be estimated under the constraint that all α_is equal a constant times the sum of 1 minus β_i. Obviously the constrained equation cannot have more explanatory power than the unconstrained equation. However, if it has less at a statistically significant level, it would be strong evidence for rejecting both the standard and zero-beta forms of the CAPM. Gibbons performs this test using the methodology of seemingly unrelated regression assessing the gamma term constant through time and does a likelihood ratio test on the difference in explanatory power between the constrained and

unconstrained regression. Defining the market as an equally weighted portfolio of New York Stock Exchange stocks, Gibbons rejects both the standard form and the zero-beta form of the CAPM.

Using a slightly different (Lagrange multiplier) test and an extended definition of the market portfolio including corporate bonds, government bonds, Treasury bills, home furnishings, residential real estate, and automobiles, Stambaugh (1982) takes a similar approach to Gibbons in examining the CAPM. However, he uses a different statistical test (a Lagrangian multiplier test rather than a likelihood ratio test). Stambaugh claims that his test is more powerful for samples of the size studied by both authors, and based on his test, he reaches very different conclusions than does Gibbons. Stambaugh finds strong support for the zero-beta form of the CAPM and evidence against the standard form. Furthermore, Stambaugh performs these tests using several samples.

The maximum likelihood approach suggested in the papers of Gibbons and of Stambaugh does not necessarily require the market model assumption [Equation (15.5)]. The approach can easily generalize to cases where the residuals in Equation (15.5) are heteroskedastic and are correlated across securities. McElroy and Burmeister (1988) observe that there exists off-the-shelf technology to estimate these models. This technology does not require that we observe the true market portfolio and can be generalized to consider multiple factor representations of the security-generating process.

But what of the simple Fama and MacBeth procedure? Brown and Weinstein (1983) observe that Equation (15.6) is in the form of a bilinear model for which there is a very large established literature in statistics. In fact, Equation (15.6) corresponds to the very earliest single-factor version of the factor analysis model considered in the next chapter. Spearman (1904) pioneered the Fama and MacBeth time series/cross-sectional regression approach, which he proposed as an algorithm to estimate models of this nature. Brown and Weinstein observe that given maximum likelihood estimates of time-varying gammas, simple time series regressions define betas, and given maximum likelihood betas, gammas are given by generalized least squares cross-sectional regressions. Thus the Fama and MacBeth procedure can be considered a first step toward a maximum likelihood solution. By comparing the Fama and MacBeth estimators with maximum likelihood estimators, Shanken (1982) argues that the procedure overstates the precision of the estimate of gamma because, as Miller and Scholes point out, the betas are measured with error. He is able to derive a simple adjustment formula to the Fama and MacBeth estimates that represents a full and complete correction for the measurement error problem identified by Miller and Scholes. For the overall period from 1935 to 1968, the Fama and MacBeth estimated prices of risk range from 0.72% per month to 1.14% per month. Shanken's adjustment for the Miller and Scholes errors-in-variables problem accounts for only a maximum of 0.035% per month. Furthermore, generalizing Shanken's work, Jagannathan and Wang (1996) show that the Fama and MacBeth procedure does not, in fact, overestimate the precision of gamma estimates. A net consequence of all of this current research is that the simple Fama and MacBeth procedure addresses the Miller and Scholes critique and is an essential tool used by all practical asset pricing empiricists today.

A further intuition is gained by the zero-correlation result of Fama and MacBeth. The equilibrium model in fact demands that the residuals do not covary with all other parameters and explanatory variables in the asset pricing relation. This result does not depend on any normal distribution result. For all asset pricing model variants, we can (perhaps with some difficulty) write down expressions for these covariances. By setting these expressions to zero and solving for the parameters, we can obtain estimates that are robust to nonnormality. Hansen and Singleton (1982) define the statistical properties of this Generalized Method of Moments procedure. The intuition behind these results is at the

same time both simple and very powerful. In most practical circumstances, there are many more covariances than there are parameters to be estimated. By the central limit theorem, the sample covariances are asymptotically normal, and using least squares to fit the parameters to these covariances, we obtain estimates that are also asymptotically normal. Moreover, the extent to which the parameters set the covariances to zero provides a simple chi square test based on the number of covariances in excess of the number of betas and gammas to be estimated. Finally, because the iterative maximum likelihood procedure proposed by McElroy and Burmeister (1988) automatically satisfies the zero-correlation equations, we can assert that it too is robust to nonnormality, another issue raised by Miller and Scholes.

With access to the powerful tools provided by maximum likelihood methods and Generalized Method of Moments, we can relax almost all of the limiting assumptions that underlie the empirical asset pricing model. Jagannathan and Wang (1996) argue that the CAPM applies period by period but that the beta and gamma coefficients vary with economic conditions.[8] Furthermore, a true index of the return on market wealth must of necessity include the return on human capital. As of the end of 1986, the total market value of equities held by the households category was 80% of GNP, while mortgages, consumer credit, and bank loans to the household sector add up to the same fraction of GNP. Jagannathan and Wang propose a proxy for the return on market wealth given as the return on the value-weighted equity portfolio and an index of the change in labor income. Their results are quite striking. Applying the standard Fama and MacBeth procedure to 100 portfolios organized by prior beta and size and including the logarthim of prior equity market value along with beta as explanatory variables, beta has no effect ($t = -0.94$), while size is very important in explaining the cross-sectional dispersion of returns ($t = -2.30$). Similar results are obtained when we augment the definition of the market return to include labor income. However, when we allow gamma to vary systematically with the yield spread between BAA- and AAA-rated bonds, the results are reversed. Now the market gamma is highly significant, and the size effect drops out completely ($t = -1.45$). These results are obtained under the zero-beta representation of the model. Even allowing for time variation in gamma, the data convincingly reject the hypothesis that the zero-beta rate is equal to the risk-free rate. Measuring returns in excess of the Treasury bill rate, the intercept is positive and significant ($t = 3.58$). The weight of the evidence continues to favor the zero-beta version of the CAPM.

TESTING SOME ALTERNATIVE FORMS OF THE CAPM MODEL

It is difficult to state that any form of the CAPM is right or wrong. In fact, in the next section of this chapter we will see that there are additional problems we have not as yet faced. Although it may be impossible to accept or reject a model as correct for all purposes, it may be possible to say that one form of a model works better for a specific purpose or explains historical returns better than another form of a model. The nonstandard forms of the CAPM described in Chapter 14 have not been subject to the intense investigation that has been performed on the more standard CAPMs. However, there are two models that have been investigated in some detail: the posttax form of the CAPM and the consumption-based CAPM. We discuss each briefly.

[8]Jagannathan and Wang were not the first to consider conditional tests of the asset pricing model. Gibbons and Ferson (1985) and Ferson, Kandel, and Stambaugh (1987) pioneered the approach to estimating the asset pricing model with time-varying parameters.

TESTING THE POSTTAX FORM OF THE CAPM MODEL

Although a great deal of attention has been paid to tests of the zero-beta (two-factor) CAPM model, almost no testing has been done on the other forms of general equilibrium models described in the previous chapter. The one exception to this is tax-adjusted versions of the general equilibrium model. Black and Scholes (1974) have tested a form of the CAPM that includes a dividend term and concluded that dividends do not affect the equilibrium relationship. Because a dividend term is present in the posttax CAPM, this would seem to indicate that a pretax CAPM is more descriptive of equilibrium returns. However, subsequently, Litzenberger and Ramaswamy (1979) have found strong, positive support for dividends affecting equilibrium prices. Their results differ from Black and Scholes at least in part because while Black and Scholes assumed that dividends were received in equal amounts each month, Litzenberger and Ramaswamy formulated their tests so that dividends were assumed to be received in the month in which they could reasonably be expected to occur.[9] They tested a model of the form

$$R_{it} - R_{Ft} = \gamma_0 + \gamma_1 \beta_{it} + \gamma_2 (\delta_{it} - R_{Ft}) + e_{it}$$

where δ_{it} is the dividend divided by price for stock i in month t. This model appears like a test of the two-factor model with the addition of a new term involving dividend yields. The form of this new term is consistent with the posttax model presented in Chapter 14 with γ_2 interpreted as τ.[10]

When Litzenberger and Ramaswamy tested this model using maximum likelihood estimates on monthly data, they found the following results for the period 1936–1977:[11]

$$R_{it} - R_{Ft} = 0.0063 + 0.0421\beta_{it} + 0.236(\delta_{it} - R_{Ft})$$
$$(2.63) \qquad (1.86) \qquad (8.62)$$
$$t\text{-statistics in parentheses}$$

The key point to note from this analysis is that the dividend term is positive and statistically significant. Furthermore, it is obvious that the dividend term is of economic significance. This term indicates that for every $1 of dividends paid, stock investors require 23.6¢ in extra return. The model also allows us to infer the effective tax rates for determining equilibrium in the market. Recall that γ_2 is equal to τ.

In Chapter 14 we demonstrated that τ was equal to an average of τ_i:

$$\tau_i = \frac{t_{di} - t_{gi}}{1 - t_{gi}}$$

where

t_{di} = tax rate paid on dividend income

t_{gi} = tax rate paid on capital gain income

---

[9]Other differences are that Litzenberger and Ramaswamy based the form of their dividend term on the general equilibrium equation. Black and Scholes simply added a dividend term to the standard CAPM. In addition, Litzenberger and Ramaswamy used maximum likelihood methods for estimating their equation, rather than relying on the portfolio grouping techniques of Black and Scholes.

[10]Variable τ is related to tax rates, as explained in text below.

[11]Litzenberger and Ramaswamy estimate this equation for six subperiods during the 1936–1977 time span. In each subperiod the dividend yield term has positive signs. It is statistically significant in five of the six periods. This is the best behaved of the three coefficients, as each of the other coefficients has the wrong sign in two subperiods and is statistically significant in only one or two of the subperiods.

The assumption behind this derivation was that capital gains taxes as well as ordinary income taxes were paid at the end of each period (e.g., year). Litzenberger and Ramaswamy developed an analogous model under the assumption that capital gains taxes are postponed indefinitely and are essentially equal to zero. Under this assumption, t_{gi} equals zero and τ equals an average of t_{di}. The truth probably lies somewhere between these two extremes. Using their estimate of τ, the effective income tax rate lies in the following range:[12]

$$0.236 \leq t_{di} \leq 0.382$$

They also tested for and found evidence supporting the presence of a clientele effect. That is, stockholders in high tax brackets tended to hold stocks with low dividend yields, while investors in low tax brackets tended to hold stocks with high dividend yields. These results are consistent with the findings of Elton and Gruber (1970).

Testing the Consumption-Based CAPM (CCAPM)

In an effort to improve upon the empirical specification of the static CAPM, a series of papers have formulated tests of the consumption-based CAPM.[13] One of the most comprehensive sets of tests is found in a paper by Breeden, Gibbons, and Litzenberger (1989). The form of the model they test has been examined in Chapter 14 and is briefly summarized here. Returns are assumed to be generated by the following process:

$$R_{it} = \alpha_i + \beta_i C_t + e_{it}$$

where by assumption

$E(e_{it}) = 0$ and the covariance between e_{it} and C_t is zero or $E(e_{it}C_t) = 0$

Under this model

$$\beta_i = \frac{\text{Cov}(R_{it}, C_t)}{\text{Var}(C_t)}$$

and the equilibrium return for any security is given by

$$\overline{R}_i = \overline{R}_z + \gamma_1 \beta_i$$

where

C_t = the rate of growth in per capita consumption at time t

γ_1 = the market price of consumption risk (beta)

As pointed out earlier, this set of equations is analogous to the equations for the zero-beta form of the CAPM, with the return on the market portfolio replaced by the rate of growth in consumption between two points in time.

---

[12]The lower estimate is of course their coefficient on the dividend term. The higher estimate is obtained by setting

$$\frac{t_{di} - t_{gi}}{1 - t_{gi}} = 0.236 \qquad \text{and} \qquad t_{gi} = (1/2)t_{di}$$

[13]See Breeden (1979, 1980), Breeden and Litzenberger (1978), and Breeden, Gibbons, and Litzenberger (1989).

Testing the consumption capital asset pricing models has many econometric problems in common with testing the zero-beta form of the standard capital asset pricing model. The major problem both have in common is identifying the variable that drives return (in this case the growth rate in per capita consumption).

Breeden, Gibbons, and Litzenberger (1989) have recognized and attempted to solve four types of problems that arise in measuring the rate of growth in per capita consumption. These measurement problems stem from the fact that

1. any estimate of consumption contains sampling error.
2. statistics are reported on expenditures, not on consumption.
3. total expenditures over some period of time (a month or a quarter) are reported rather than expenditures at a point in time.
4. after 1958 monthly numbers are reported, but only quarterly expenditures are reported for the period prior to 1958.[14]

Breeden, Gibbons, and Litzenberger show that if errors in measuring consumption are random and uncorrelated with economic variables, the estimate of the price of risk (γ_1) will be upward biased, but their tests of the significance of the model will not be biased. Breeden, Gibbons, and Litzenberger deal with the second problem by assuming that expenditures on nondurable goods plus services act as a good proxy for consumption. They ignore any consumption flow from durable goods and any pattern in the storage of nondurables.

The third problem is more difficult to solve. Because consumption expenditures are reported for a period of time rather than at a point in time, expenditures are averaged. Estimated betas on averaged consumption are less than they would be if consumption were reported at a point in time. They estimate the size of this difference and rescale the growth in consumption so that the betas are as if consumption were reported at a point in time.[15]

Having shown this, the authors are left with the last remaining problem, the unavailability of monthly data. Breeden (1979) has shown that the CAPM holds when the growth in aggregate per capita consumption is replaced with the rate of return on a portfolio of assets that has maximum correlation with the appropriate consumption series. By designing such a portfolio using quarterly data, the authors can then proceed to test the consumption CAPM using monthly observations on this portfolio (called the consumption portfolio). Breeden, Gibbons, and Litzenberger employ data from 1929–1982 to find the consumption portfolio (MCP), which has maximum correlation with consumption. The portfolio is formed from among return series on each of 13 industries plus return series for U.S. Treasury bills, long-term government bonds, long-term corporate bonds, and a junk bond premium. The composition of the MCP portfolio is assumed to be the same over the entire time period, 1929–1982. The MCP is a portfolio of stocks and debt instruments that is clearly related to but different from many of the proxies that have been used for the market portfolio. For the period of study, 1929–1982, the correlation between the MCP and the CRSP value-weighted index is 0.67.

---

[14]The authors use expenditures on nondurable goods and services based on national income accounting. The Commerce Department statistics on average U.S. population are used to obtain per capita statistics.

[15]In the next paragraph we discuss construction of a matching portfolio. Betas are needed to construct this matching portfolio. The adjustment is determined as follows. Breeden, Gibbons, and Litzenberger show analytically that the variance in this smoothed series should be equal to two-thirds of the variance of the unsmoothed series, that the covariance of the smoothed series with spot quarterly returns on securities should be equal to one-half of the spot covariance with the unsmoothed series, and thus that the smoothed beta of return with consumption should be equal to three-quarters of the unsmoothed beta. This analysis is used to adjust the growth in consumption so that betas will be the appropriate size.

While the consumption CAPM was originally motivated by a desire to improve upon the empirical performance of the standard CAPM, Hansen and Singleton (1982, 1983) show that the consumption-based model is convincingly rejected by the data; in fact, the standard CAPM better explains the cross section of asset returns than does the consumption CAPM (Campbell 1996; Cochrane 1996). Just as Jagannathan and Wang are able to resurrect the CAPM by considering the fact that the beta and gamma parameters may depend on general economic conditions, Lettau and Ludvigson (2001) condition their consumption CAPM on a measure of the aggregate consumption to wealth ratio. Not only does the consumption CAPM now fit the data quite well, but also this model outperforms the standard static CAPM model in being able to explain the value and size premia otherwise unaccounted for by the simple model.

SOME RESERVATIONS ABOUT TRADITIONAL TESTS OF GENERAL EQUILIBRIUM RELATIONSHIPS AND SOME NEW RESEARCH

In this chapter we have reviewed some of the classic tests of general equilibrium relationships. These tests were intended to validate the theories we have described in the previous two chapters. Roll (1985) has argued that general equilibrium models of the form of the CAPM are not amenable to testing or, at least, that the tests performed so far provide little evidence in support of, or against, CAPM. Roll raised some legitimate questions, and his arguments are well worth reviewing.

Perhaps the easiest way to understand Roll's case is to start with his proof that if *any* ex post mean variance efficient portfolio is selected as the market portfolio and betas are computed using this as the market proxy, then the equation

$$\overline{R}_i = \overline{R}_{ZP} + \beta_{iP}\left(\overline{R}_P - \overline{R}_{ZP}\right)$$

must hold.[16] In fact, it is a tautology that has nothing to do with the way equilibrium is set in the capital markets or with investor's attitude toward risk.

Proof Return to Equation (13.4). Assume a riskless asset exists with a return R_F. Then

$$\lambda\left(X_1\sigma_{1k} + X_2\sigma_{2k} + \cdots + X_k\sigma_k^2 + \cdots + X_N\sigma_{kN}\right) = \overline{R}_k - R_F$$

If all X_is stand for the proportion of stock i in portfolio P, we can write this expression as

$$\lambda\sigma_{kP} = \overline{R}_k - R_F \tag{15.7}$$

Because this expression must hold for each security in portfolio P, it must also hold for portfolio P itself, or

$$\lambda\sigma_P^2 = \overline{R}_P - R_F$$

Solving for λ, substituting in Equation (15.5), and rearranging,

$$\overline{R}_k = R_F + \frac{\sigma_{kP}}{\sigma_P^2}\left(\overline{R}_P - R_F\right)$$

[16]In this expression P is the proxy for the market portfolio, $\overline{R}_P$ is the expected return on the proxy for the market portfolio, β_{iP} is the beta for security i with the proxy market portfolio, and R_{ZP} is the minimum-variance portfolio that has a zero beta with the market proxy portfolio.

Recognizing that $(\sigma_{kP}/s_P^2) = \beta_{kP}$, we can write this as

$$\overline{R}_k = R_F + \beta_{kP}\left(\overline{R}_P - R_F\right) \tag{15.8}$$

Now, as we did in Chapter 14, assume that lending and borrowing cannot take place at the riskless rate R_F. However, as we have seen, an infinite number of portfolios will exist that have the return R_F. From Equation (15.8) they all must be uncorrelated with portfolio *P*. Let $\overline{R}_{ZP}$ stand for the minimum-variance portfolio that is uncorrelated with portfolio *P*. Then, because $\overline{R}_{ZP} = R_F$, Equation (15.8) can be written as

$$\overline{R}_k = \overline{R}_{ZP} + \beta_{kP}\left(\overline{R}_P - \overline{R}_{ZP}\right)$$

From the proof it follows that the return on an asset or portfolio is an *exact* linear function of beta if betas are computed using any efficient portfolio. Conversely, if the portfolio used to compute betas is not efficient, then return is not an exact linear function of beta.

A number of authors have tried to address the Roll critique by broadening the definition of the market portfolio by including measures of human capital and by extending the model to consider intertemporal consumption/investment decisions. Shanken (1986) suggests a creative response to the dilemma of not being able to observe the market portfolio. When we use one or more proxies for the true but unobserved market portfolio, the fit of the implied CAPM will obviously depend on how close the proxy or proxies are to the true market. For example, if an equally weighted market portfolio explains only half of the variance of the true market, Shanken shows that we will reject the CAPM at the 10% level even if the CAPM is true.

However, Roll appears to be making a more profound observation. Mean–variance efficiency of the market portfolio implies CAPM. The reverse implication does not necessarily follow. lf we go to the data and find that the betas we estimate explain the cross-sectional dispersion of average returns, Roll would argue that this finding is silent on the issue of whether the observed market portfolio is mean–variance efficient. In other words, tests of the linear equilibrium model have no power as tests of mean–variance efficiency. Roll's critique was a challenge to the literature to develop tests that indeed have power as tests of mean–variance efficiency.

An interesting approach to this issue was pioneered by Jim MacBeth in his University of Chicago dissertation (1975). Instead of examining the linearity of the asset pricing relation, he considered the alphas, the deviations from the asset pricing relation on an individual security or portfolio basis. These alphas could be thought of as the return to a portfolio strategy that exploits information not generally available by purchasing or selling certain securities, financing the purchase (or sale) by a short (or long) position in the market portfolio and riskless asset with the same beta. We can examine the statistical significance of this excess return by dividing the average value by the standard deviation of excess return to obtain a *t*-value.[17]

MacBeth's T^2 test is more than just another test of the CAPM. Large values of T^2 imply there are large profit opportunities by investing outside the market portfolio. Because the value is computed on the basis of publicly available information, these profit opportunities must imply some inefficiency in the market portfolio. This intuition can be made precise. Roll (1985) demonstrates that T^2 is in fact proportional to the extent to which the measured market portfolio lies within the minimum-variance frontier. Large values of MacBeth's T^2 translate directly into measures of the reduction in variance risk possible by

[17]If the market portfolio is inefficient.

moving from the measured market portfolio to a more efficient portfolio with the same mean return. For this reason, MacBeth's T^2 test is a direct test of mean–variance efficiency.

Gibbons, Ross, and Shanken (1989) consider in detail how to interpret the T^2 measure.[18] They show that T^2 can be expressed as

$$\frac{\theta_p^2 - \theta_m^2}{\theta_m^2}$$

where θ_p^2 is the square of the implied Sharpe ratio for the optimum mix of assets or portfolios being analyzed and θ_m^2 is the squared Sharpe ratio on the market portfolio. The Sharpe ratio is excess return (over the risk-free rate) divided by the standard deviation of returns.

In other words, if the T^2 measure is large, this means that there exists another portfolio P which has a Sharpe ratio θ_p^2 superior to that of the market portfolio. This is just a fancy way of saying that the market portfolio cannot be mean–variance efficient.

Gibbons, Ross, and Shanken replicate the study of Black, Jensen, and Scholes (1972) using this new technology. On the basis of 10 beta-sorted portfolios over the period considered by the earlier work, the implied optimal Sharpe ratio θ_p is 0.227 while for the same period the CRSP Equally Weighted Market Index as a proxy for the market portfolio has θ_m equal to 0.166. They conclude that there is no evidence of a significant deviation from mean–variance efficiency in the Black, Jensen, and Scholes database. A similar result follows with size-sorted portfolios, but only when the month of January is excluded from the analysis. When they include the month of January, the Sharpe ratios are significantly different from one another. It would appear that including size and calendar descriptors into the analysis does damage to the mean–variance efficiency hypothesis. However, it should be noted that this test is constructed on the basis of the unconditional mean and covariance matrix of returns. More recent work by Jagannathan and Wang (1996) and Lettau and Ludvigson (2001) suggests that we should be looking at a more forward-looking concept of mean–variance efficiency that is conditional in that it accounts for changes in economic conditions. However, Lewellen and Nagel (2006) show that while betas and risk premia do change with economic conditions, the necessary changes would have to be implausibly large to explain observed return regularities. Conditional alphas are large and significant in violation of the conditional CAPM.

CONCLUSION

If we reexamine the tests in this chapter, not as tests of the CAPM, but as inputs to the portfolio process, do we gain useful information? We would argue that we do. The fact that

[18]If the market portfolio were inefficient, then we should observe a large number of significant *t*s; the only issue is how to aggregate these *t*s for an overall test. Because some are positive and some are negative, we need to square them. With a slight modification to account for correlation among residuals, the test is the significance of the sum of the squared *t*s called T^2. Each square is asymptotically chi-square distributed, and if independent, the sum would be distributed as chi-square with degrees of freedom equal to the number of securities or portfolios we are considering. But the alphas are correlated across securities and portfolios. To account for this fact, instead of normalizing the squared excess returns by the reciprocal of excess return variance, we normalize using the inverse covariance matrix of excess returns. This statistic corresponds to Hotelling's T^2 measure

$$T^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \omega_{ij}$$

where ω_{ij} is the i, jth element of Ω, the inverse covariance matrix of excess returns.

return and risk appear to be linearly related for securities and portfolios over long periods of time, when risk is defined as systematic risk, is important. The same can be said for the fact that return is not related to residual risk. Even if these statements do not constitute tests of the CAPM, they have important implications for behavior. Investors are not rewarded for taking nonmarket risk, but they are rewarded for bearing added market risk. These statements seem to hold under alternative methods of calculating systematic risk. Furthermore, they seem to hold even more firmly when systematic risk is calculated using a value-weighted, rather than equally weighted, market proxy. The fair game nature of the model is also important. Not only does the model seem to hold over long periods of time, but intertemporal deviations from the model cannot be used to make an extra return. In summary, while the empirical work is not fully a satisfactory test of the CAPM, it produces results that are consistent with what one would expect from a test of the CAPM. Furthermore, these results are produced with respect to observables variables (market proxies). While we should continue to search for true tests of the CAPM, we can, with some care, proceed on the basis of the results produced by tests of observable, but not optimum, phenomena.

There is another direction that testing can take. In this chapter we have attempted to see whether we could prove the CAPM was "true" or not. A very practical question to ask is if there is another model of asset prices that gives us added insight into capital markets. Recent evidence suggests that CAPM is descriptive of the data once we consider the model in a forward looking sense that adjusts for changes in economic conditions. This has important implications not only for asset pricing but also for practical applications to optimal portfolio selection. In the next chapter we describe a competing paradigm for describing asset prices. We then examine tests of that model to see if, in fact, it allows us to gain new insights into portfolio management and helps to explain what happens in capital markets.

QUESTIONS AND PROBLEMS

1. We have sometimes heard investment managers say, "I followed that (expletive deleted) theory and bought high-beta stocks last year and they did worse than low-beta stocks. That theory is (expletive deleted)." Is this a valid test, and is this empirical evidence inconsistent with the theory?

2. A new theory has been proposed. The expected percentage increase in alcoholism in each city is equal to the rate of change in the price of gold plus the product of two terms. The first is the covariance of the percentage change in alcoholism in the city with the percentage change in professors' salaries divided by the variance of the percentage change in professors' salaries. The second term is the percentage change in professors' salaries minus the percentage increase in gold. How would you test this proposition?

3. Show that if the market portfolio is not an efficient portfolio, then

$$\overline{R}_i = \overline{R}_Z + \beta_i\left(\overline{R}_M - \overline{R}_Z\right)$$

cannot in general hold.

4. Explain how you might use general equilibrium theory to evaluate the performance of one or more common-stocks managers.

5. Assume the posttax CAPM holds but the Sharpe–Lintner model is tested. What would you expect the empirical results to look like?

BIBLIOGRAPHY

1. Alder, Michael. "On the Risk-Return Trade-Off in the Valuation of Assets," *Journal of Financial and Quantitative Analysis*, **IV,** No. 4 (Dec. 1969), pp. 492–512.
2. Bar-Yosef, Sasson, and Kolodny, Richard. "Dividend Policy and Capital Market Theory," *Review of Economics and Statistics*, **LVIII,** No. 2 (May 1976), pp. 181–190.
3. Belkaoui, Ahmed. "Canadian Evidence of Heteroscedasticity in the Market Model," *Journal of Finance*, **XII,** No. 4 (Sept. 1977), pp. 1320–1324.
4. Best, Michael J., and Grauer, Robert R. "Capital Asset Pricing Compatible with Observed Market Value Weights," *The Journal of Finance*, **40,** No. 1 (March 1985), pp. 85–104.
5. Black, F., and Scholes, M. "The Effects of Dividend Yield and Dividend Policy on Common Stock Prices and Returns," *Journal of Financial Economics*, **1** (1974), pp. 1–22.
6. Black, F., Jensen, M. C., and Scholes, M. "The Capital Asset Pricing Model: Some Empirical Tests," in M. C. Jensen (ed.), *Studies in the Theory of Capital Markets* (New York: Praeger, 1972).
7. Blume, Marshall, and Friend, Irwin. "A New Look at the Capital Asset Pricing Model," *Journal of Finance*, **VIII,** No. 1 (March 1973), pp. 19–33.
8. —. "Risk, Investment Strategy, and the Long-Run Rates of Return," *Review of Economics and Statistics*, **LVI,** No. 3 (Aug. 1974), pp. 259–269.
9. Blume, Marshall, and Husic, Frank. "Price, Beta, and Exchange Listings," *Journal of Finance*, **VIII,** No. 2 (May 1973), pp. 283–299.
10. Breeden, D. "An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities," *Journal of Financial Economics*, **7** (1979), pp. 265–296.
11. —. "Consumption Risk in Futures Markets," *Journal of Finance*, **35** (1980), pp. 503–520.
12. Breeden, D., and Litzenberger, R. "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, **51** (1978), pp. 621–651.
13. Breeden, D., Gibbons, M., and Litzenberger, R. "Empirical Tests of the Consumption-Oriented CAPM," *Journal of Finance*, **44** (1989), pp. 231–262.
14. Brown, David P., and Gibbons, Michael R. "A Simple Econometric Approach for Utility-Based Asset Pricing Models," *The Journal of Finance*, **40,** No. 2 (June 1985), pp. 359–382.
15. Brown, Stephen J., and Weinstein, Mark I. "A New Approach to Testing Asset Pricing Models: The Bilinear Paradigm," *The Journal of Finance*, **38,** No. 3 (June 1983), pp. 711–744.
16. Campbell, John. "Understanding Risk and Return," *Journal of Political Economy*, **104** (1996), pp. 298–345.
17. Campbell, John, Lo, Andrew, and MacKinlay, A. Craig. *The Econometrics of Financial Markets* (Princeton, NJ: Princeton University Press, 1997).
18. Chamberlain, G., and Rothschild, M. "Arbitrage, Factor Structure, and Mean–Variance Analysis on Large Asset Markets," *Econometrica*, **51** (1983), pp. 1281–1304.
19. Chen, N., Roll, R., and Ross, S. "Economic Forces and the Stock Market," *Journal of Business*, **59** (1986), pp. 386–403.
20. Clarkson, Pete, Guedes, José, and Thompson, Rex. "On the Diversification, Observability, and Measurement of Estimation Risk," *Journal of Financial and Quantitative Analysis*, **31,** No. 1 (March 1996), pp. 69–84.
21. Cochrane, John H. "A Cross-Sectional Test of an Investment-Based Asset Pricing Model." *Journal of Political Economy*, **104** (1996), pp. 572–621.
22. Connor, G. "A Unified Beta Pricing Theory," *Journal of Economic Theory*, **34** (1984), pp. 13–31.
23. Connor, G., and Korajczyk, R. "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, **15** (1986), pp. 373–394.
24. Cornell, B. "The Consumption Based Asset Pricing Model: A Note on Potential Tests and Applications," *Journal of Financial Economics*, **9** (1981), pp. 103–108.
25. Dhrymes, Phoebus, Friend, Irwin, and Gultekin, Bulent. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory," *The Journal of Finance*, **39** (June 1984), pp. 323–346.
26. Douglas, George. *Risk in the Equity Markets: An Empirical Appraisal of Market Efficiency* (Ann Arbor, MI: University Microfilms, Inc., 1968).

27. Dybvig, Phillip H. "An Explicit Bound on Deviations from APT Pricing in a Finite Economy," *Journal of Financial Economics*, **12** (1983), pp. 483–496.
28. Dybvig, P., and Ross, S. "Yes, the APT Is Testable," *Journal of Finance*, **40** (1985), pp. 1173–1188.
29. Elton, Edwin J. "Presidential Address: Expected Return, Realized Return and Asset Pricing Tests," *Journal of Finance*, **54** (Aug. 1999), pp. 1199–1220.
30. Elton, Edwin J., and Gruber, Martin J. "Marginal Stockholder Tax Rates and the Clientele Effect," *Review of Economics and Statistics*, **52** (1970), pp. 68–74.
31. Eubank, Arthur. "Risk-Return Contrasts: NYSE, AMEX, and OTL," *Journal of Portfolio Management*, **3,** No. 4 (Summer 1977), pp. 25–30.
32. Fama, Eugene, and MacBeth, J. "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, **71** (May/June 1973), pp. 607–636.
33. ——. "Tests of the Multiperiod Two-Parameter Model," *Journal of Financial Economics*, **1,** No. 1 (May 1974), pp. 43–66.
34. Fama, E., MacBeth, J., and Schwert, G. "Asset Returns and Inflation," *Journal of Financial Economics*, **5** (1977), pp. 115–146.
35. ——. "Inflation, Interest and Relative Prices," *Journal of Business*, **52** (1979), pp. 183–209.
36. Fama, Eugene, and French, Kenneth. "The Cross-Section of Expected Stock Returns," *Journal of Finance*, **67** (1992), pp. 427–465.
37. Ferson, W. "Expected Real Interest Rates and Consumption in Efficient Financial Markets: Empirical Tests," *Journal of Financial and Quantitative Analysis*, **18** (1983), pp. 477–498.
38. Ferson, W., Kandel, S., and Stambaugh, R. "Tests of Asset Pricing with Time-Varying Expected Risk Premiums and Market Betas," *Journal of Finance*, **42** (1987), pp. 201–220.
39. Foster, George. "Asset Pricing Models: Further Tests," *Journal of Financial and Quantitative Analysis*, **XIII,** No. 1 (March 1978), pp. 39–53.
40. Friend, Irwin, Westerfield, Randolf, and Granito, Michael. "New Evidence on the Capital Asset Pricing Model," *Journal of Finance*, **XII,** No. 3 (June 1978), pp. 903–917.
41. Gentry, James, and Pike, John. "An Empirical Study of the Risk-Return Hypothesis Using Common Stock Portfolios of Life Insurance Companies," *Journal of Financial and Quantitative Analysis*, **V,** No. 2 (May 1970), pp. 179–185.
42. Gibbons, Michael R. "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics*, **X,** No. 1 (March 1982), pp. 3–28.
43. Gibbons, Michael R., and Ferson, Wayne. "Testing Asset Pricing Models with Changing Expectations and an Unobservable Market Portfolio," *Journal of Financial Economics*, **XIV,** No. 2 (June 1985), pp. 217–236.
44. Gibbons, Michael, Ross, Stephen, and Shanken, Jay. "A Test of the Efficiency of a Given Portfolio," *Econometrica*, **57** (1989), pp. 1121–1152.
45. Grinblatt, Mark, and Titman, Sheridan. "Factor Pricing in a Finite Economy," *Journal of Financial Economics*, **12** (1983), pp. 497–507.
46. Grinblatt, Mark, and Titman, Sheridan. "The Relation between Mean–Variance Efficiency and Arbitrage," *The Journal of Business*, **60,** No. 1 (Jan. 1987), pp. 97–112.
47. Grossman, S., and Shiller, R. "Consumption Correlatedness and Risk Measurement in Economies with Non-traded Assets and Heterogeneous Information," *Journal of Financial Economics*, **10** (1982), pp. 195–210.
48. Grossman, S., Melino, A., and Shiller, R. "Estimating the Continuous-Time Consumption-Based Asset-Pricing Model," *Journal of Business and Economic Statistics*, **5** (1987), pp. 315–328.
49. Hall, R. "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, **86** (1978), pp. 971–987.
50. Hansen, L., and Singleton, K. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica*, **50** (1982), pp. 1269–1286.
51. ——. "Stochastic Consumption, Risk Aversion, and the Temporary Behavior of Asset Returns," *Journal of Political Economy*, **91** (1983), pp. 249–265.
52. Ibbotson, Roger, and Sinquefield, Rex. *Stocks, Bonds, Bills and Inflation: The Past and the Future* (Charlottesville, VA: Financial Analysts Research Foundation, 1982).

53. Ingersoll, Jonathan E., Jr. "Some Results in the Theory of Arbitrage Pricing," *Journal of Finance*, **39** (1984), pp. 1021–1039.
54. Jagannathan, Ravi, and Wang, Zhenyu. "The Conditional CAPM and the Cross-Section of Expected Returns," *Journal of Finance*, **51** (1996), pp. 3–53.
55. Jobson, J., and Korkie, B. "Estimation for Markowitz Efficient Portfolios," *Journal of the American Statistical Association*, **75** (1980), pp. 544–554.
56. ——. "Potential Performance Tests of Portfolio Efficiency," *Journal of Financial Economics*, **10** (1982), pp. 433–466.
57. Kandel, S. "On the Exclusion of Assets from Tests of the Mean–Variance Efficiency of the Market Portfolio," *Journal of Finance*, **39** (1984), pp. 63–75.
58. ——. "The Likelihood Ratio Test Statistic of Mean–Variance Efficiency without a Riskless Asset," *Journal of Financial Economics*, **13** (1984), pp. 575–592.
59. Kandel, Shmuel. "The Geometry of the Maximum Likelihood Estimator of the Zero-Beta Return," *Journal of Finance*, **41,** No. 2 (June 1986), pp. 339–346.
60. Kandel, S., and Stambaugh, R. "On Correlations and the Sensitivity of Inferences about Mean–Variance Efficiency," *Journal of Financial Economics*, **18** (1987), pp. 61–80.
61. Keim, D. "Size Related Anomalies and Stock Return Seasonality: Further Empirical Evidence," *Journal of Financial Economics*, **12** (1983), pp. 13–32.
62. Lau, Sheila, Quay, Stuart, and Ramsey, Carl. "The Tokyo Stock Exchange and the Capital Asset Pricing Model," *Journal of Finance*, **IX,** No. 2 (May 1974), pp. 507–514.
63. Lehmann, Bruce N., and Modest, David M. "The Empirical Foundations of the Arbitrage Pricing Theory," *Journal of Financial Economics*, **21,** No. 2 (Sept. 1988), pp. 213–254.
64. Lettau, Martin, and Ludvigson, Sydney. "Consumption Aggregate Wealth and Expected Stock Returns," *Journal of Finance*, **LVI,** No. 3 (2001), pp. 815–849.
65. Lewellen, Jonathan, and Nagel, Sefan. "The Conditional CAPM Does Not Explain Asset-Pricing Anomalies," *Journal of Financial Economies*, **82,** No. 2 (2006), pp. 289–314.
66. Lewellen, Jonathan, Nagel, Stefan, and Shanken, Jay. "A Sceptical Appraisal of Asset Pricing Tests," *Journal of Financial Economics*, **96,** No. 2 (2010), pp. 175–194.
67. Litzenberger, R. H., and Budd, A. P. "Secular Trends in Risk Premiums," *Journal of Finance*, **VII,** No. 3 (June 1972), pp. 857–864.
68. Litzenberger, R. H., and Ramaswamy, K. "The Effect of Personal Taxes and Dividends on Capital Asset Prices: Theory and Empirical Evidence," *Journal of Financial Economics*, **7,** No. 2 (June 1979), pp. 163–195.
69. MacBeth, J. D. "Tests of Two Parameter Models of Capital Market Equilibrium," Ph.D. dissertation, Graduate School of Business, University of Chicago (1975).
70. McElroy, M., and Burmeister, E. "Arbitrage Pricing Theory as a Restricted Nonlinear Regression Model," *Journal of Business and Economic Statistics* **6** (1988), pp. 29–42.
71. Merton, Robert C. "An Intertemporal Capital Asset Pricing Model," *Econometrica*, **41** (1973), pp. 867–887.
72. Miller, M. H., and Scholes, M. "Rates of Return in Relation to Risk: A Re-examination of Some Recent Findings," in M. Jensen, (ed.), *Studies in the Theory of Capital Markets* (New York: Praeger, 1972).
73. Morgan, I. G. "Prediction of Return with the Minimum–Variance Zero-Beta Portfolio," *Journal of Financial Economics*, **2,** No. 4 (Dec. 1975), pp. 361–376.
74. Roll, Richard. "A Critique of the Asset Pricing Theory's Tests; Part I: On Past and Potential Testability of the Theory," *Journal of Financial Economics*, **4,** No. 2 (March 1977), pp. 129–176.
75. ——. "Orthogonal Portfolios," *Journal of Financial and Quantitative Analysis*, **XV,** No. 5 (Dec. 1980), pp. 1005–1024.
76. Roll, R. "A Note on the Geometry of Shanken's CSR T2 Test for Mean–Variance Efficiency," *Journal of Financial Economics*, **14** (1985), pp. 349–358.
77. Roll, Richard, and Ross, Stephen. "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance*, **35,** No. 5 (Dec. 1980), pp. 1073–1105.
78. Roll, Richard, and Ross, Stephen. "On the Cross-Sectional Relation between Expected Returns and Betas," *Journal of Finance*, **49,** No. 1 (1994), pp. 101–121.

79. Rubinstein, M. "The Valuation of Uncertain Income Streams and the Pricing of Options," *Bell Journal of Economics and Management Science*, **7** (1976), pp. 407–425.
80. Scholes, M., and Williams, J. "Estimating Betas from Nonsynchronous Data," *Journal of Financial Economics*, **5** (1977), pp. 309–327.
81. Shanken, J. "An Asymptotic Analysis of the Traditional Risk-Return Model," unpublished manuscript, School of Business Administration, University of California, Berkeley (1982).
82. ——. "Multivariate Tests of the Zero-Beta CAPM," *Journal of Financial Economics*, **14,** No. 3 (Sept. 1985), pp. 327–348.
83. ——. "Multi-Beta CAPM or Equilibrium-APT? A Reply," *Journal of Finance*, **40,** No. 4 (1985a), pp. 1186–1189.
84. ——. "On Exclusion of Assets from Tests of the Mean–Variance Efficiency of the Market Portfolio: An Extension," *Journal of Finance*, **41,** No. 2 (1986), pp. 331–337.
85. ——. "A Posterior-Odds Ratio Approach to Testing Portfolio Efficiency," working paper, Graduate School of Management, University of Rochester, New York (1986).
86. ——. "Testing Portfolio Efficiency When the Zero-Beta Rate Is Unknown: A Note," *Journal of Finance*, **41,** No. 1 (1986), pp. 269–276.
87. ——. "Multivariate Proxies and Asset Pricing Relations," *Journal of Financial Economics*, **18,** No. 1 (1987), pp. 91–110.
88. Shanken, Jay. "The Arbitrage Pricing Theory: Is It Testable?" *Journal of Finance*, **37** (1982), pp. 1129–1140.
89. Shanken, Jay. "On the Estimation of Beta-Pricing Models," *Review of Financial Studies*, **5** (1992), pp. 1–33.
90. Sharpe, W. F. "Risk, Market Sensitivity, and Diversification," *Financial Analysts Journal*, **28,** No. 1 (Jan.–Feb. 1972), pp. 74–79.
91. Sharpe, W. F., and Cooper, G. M. "Risk-Return Class of New York Stock Exchange Common Stocks, 1931–1967," *Financial Analysts Journal*, **28,** No. 2 (March–April 1972), pp. 46–52.
92. Sharpe, W. F., and Sosin, H. "Risk, Return, and Yield: New York Stock Exchange Common Stocks, 1928–1969," *Financial Analysts Journal*, **32,** No. 2 (March–April 1976), pp. 33–42.
93. Smith, Keith. "The Effect of Intervaling on Estimating Parameters of the Capital Asset Pricing Model," *Journal of Financial and Quantitative Analysis*, **XIII,** No. 2 (June 1978), pp. 313–332.
94. Spearman, C. "'General Intelligence' Objectively Determined and Measured," *American Journal of Psychology*, **15** (1904), pp. 201–293.
95. Stambaugh, Robert F. "On the Exclusion of Assets from Tests of the Two-Parameter Model: A Sensitivity Analysis," *Journal of Financial Economics*, **X,** No. 3 (Nov. 1982), pp. 237–268.
96. Upson, Roger, and Jessup, Paul. "Risk-Return Relationships in Regional Securities Markets," *Journal of Financial and Quantitative Analysis*, **IV,** No. 5 (Jan. 1970), pp. 677–695.

16

The Arbitrage Pricing Model APT—A Multifactor Approach to Explaining Asset Prices

All of the equilibrium models discussed in Chapters 13, 14, and 15 have their basis in mean–variance analysis. All require that it is optimal for the investor to choose investments on the basis of expected return and variance. However, definitions of returns for which means and variances are calculated differ between models. For example, in the version of the capital asset pricing model (CAPM) involving taxes, investors examine means and variances of after-tax returns. As a second example, Elton and Gruber (1982) have shown that the alternative version of CAPM under conditions of uncertain inflation can be derived by assuming that investors maximize a utility function defined in terms of the mean and variance of real as compared to nominal returns. As noted in the previous chapter, there are major obstacles to testing any of these equilibrium theories.

Ross (1976, 1977) has proposed a multifactor approach to explaining the pricing of assets. Ross had developed a mechanism that, given the process that generates security returns, derives asset prices from arbitrage arguments analogous to (but more complex than) those used in the beginning of Chapter 13 to derive CAPMs. In this chapter we first present the mechanism of arbitrage pricing theory (APT). This is the derivation of equilibrium conditions given any prespecified return-generating process.

Following this, we discuss implementation of the APT. APT theory provides interesting insight into the nature of equilibrium. However, the theory is far from easy to implement. Empirical research is still in the early stages in this area. Furthermore, alternative approaches have been advocated for implementing the theory. After discussing some of those alternatives, we present an examination of whether evidence supporting APT is necessarily inconsistent with the standard form or any alternative form of the CAPM as a model of equilibrium. We close with a discussion of both applications and advantages of APT.

APT—WHAT IS IT?

Arbitrage pricing theory is a new and different approach to determining asset prices. It is based on the law of one price: two items that are the same cannot sell at different prices. The strong assumptions made about utility theory in deriving the CAPM are not necessary. In fact, the APT description of equilibrium is more general than that provided by a CAPM-type model in that pricing can be affected by influences beyond simply means and

variances. An assumption of homogeneous expectations is necessary. The assumption of investors utilizing a mean–variance framework is replaced by an assumption of the process generating security returns. APT requires that the returns on any stock be linearly related to a set of indexes, as shown in Equation (16.1),[1]

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + \cdots + b_{ij}I_j + e_i \tag{16.1}$$

where

a_i = the expected level of return for stock i if all indexes have a value of zero
I_j = the value of the jth index that impacts the return on stock i
b_{ij} = the sensitivity of stock i's return to the jth index
e_i = a random error term with mean equal to zero and variance equal to σ_{ei}^2

For the model to fully describe the process generating security returns,[2]

$$E(e_i e_j) = 0 \qquad \text{for all } i \text{ and } j \text{ where } i \neq j$$
$$E\left[e_i\left(I_j - \bar{I}_j\right)\right] = 0 \qquad \text{for all stocks and indexes}$$

If you are beginning to get the feeling that you have seen all this before, you are right. This representation is nothing more or less than the description of the multi-index model presented in Chapter 8. APT is the description of the expected returns that can be derived when returns are generated by a single- or multi-index model meeting the conditions defined before. The contribution of APT is in demonstrating how (and under what conditions) one can go from a multi-index model to a description of equilibrium.

In the following pages we demonstrate the derivation of an APT equilibrium in two different ways. The first proof stresses the economic rationale behind APT, whereas the second proof is more mathematically rigorous.

A Simple Proof of APT

We will demonstrate the expected returns that must arise from the APT with a two-index model. Suppose that the following two-index model describes returns:

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + e_i \tag{16.2}$$

Furthermore, assume that $E(e_i e_j) \approx 0$.

If an investor holds a well-diversified portfolio, residual risk will tend to go to zero and only systematic risk will matter. The only terms in the preceding equation that affect the systematic risk in a portfolio are b_{i1} and b_{i2}. Because the investor is assumed to be concerned with expected return and risk, he need be concerned with only three attributes of any portfolio (p): $\bar{R}_p$, b_{p1}, and b_{p2}.

Let us hypothesize the existence of the three widely diversified portfolios shown in the following table.

[1]The linearity assumption is not as restrictive as it might at first appear. Any of the indexes can be a nonlinear function of a variable. It could be a variable squared, the log of a variable, or any other nonlinear transformation that seems appropriate.

[2]It is convenient, though unnecessary, to assume the indexes are uncorrelated with each other. We show in Chapter 8 that a set of correlated indexes can always be converted to a set of uncorrelated indexes. The results remain the same with uncorrelated indexes, but the mathematics is more complex.

Portfolio	Expected Return	b_{i1}	b_{i2}
A	15	1.0	0.6
B	14	0.5	1.0
C	10	0.3	0.2

We know from the concepts of geometry that three points determine a plane just as two points determine a line. The equation of the plane in $\bar{R}_p$, b_{p1}, and b_{p2} space defined by these three portfolios is[3]

$$\bar{R}_i = 7.75 + 5b_{i1} + 3.75b_{i2}$$

The expected return and risk measures of any portfolio of these three portfolios are given by

$$\bar{R}_P = \sum_{i=1}^{N} X_i \bar{R}_i$$

$$b_{p1} = \sum_{i=1}^{N} X_i b_{i1}$$

$$b_{p2} = \sum_{i=1}^{N} X_i b_{i2}$$

$$\sum_{i=1}^{N} X_i = 1$$

Because a weighted combination of points on a plane (where the weights sum to one) also lies on the plane, all portfolios constructed from portfolios *A*, *B*, and *C* lie on the plane described by portfolios *A*, *B*, and *C*.[4]

What happens if we consider a new portfolio not on this plane? For example, assume a portfolio *E* exists with an expected return of 15%, a b_{i1} of 0.6, and a b_{i2} of 0.6.

Compare this with a portfolio (call it *D*) constructed by placing $\frac{1}{3}$ of the funds in portfolio *A*, $\frac{1}{3}$ in portfolio *B*, and $\frac{1}{3}$ in portfolio *C*. The b_{pj}s on this portfolio are

$$b_{p1} = \frac{1}{3}(1.0) + \frac{1}{3}(0.5) + \frac{1}{3}(0.3) = 0.6$$

$$b_{p2} = \frac{1}{3}(0.6) + \frac{1}{3}(1.0) + \frac{1}{3}(0.2) = 0.6$$

The risk for portfolio *D* is identical to the risk on portfolio *E*. The expected return on portfolio *D* is

$$\frac{1}{3}(15) + \frac{1}{3}(14) + \frac{1}{3}(10) = 13$$

[3]The reader interested in verifying this can recall that the equation of a plane can be written as $R_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2}$. By substituting in the values of R_i, b_{i1}, and b_{i2} for portfolios *A*, *B*, and *C*, we obtain three equations with three unknowns: λ_0, λ_1, and λ_2. Solving the three equations gives the values of λ_0, λ_1, and λ_2 shown in the equation in the text.

[4]The reader is encouraged to form a portfolio of portfolios *A*, *B*, and *C* with any set of X_i summing to 1. One can then see that this portfolio lies on the plane given by $\bar{R}_i = 7.75 + 5\, b_{i1} + 3.75\, b_{i2}$. One example of this is portfolio D, which is analyzed shortly in the text.

Alternatively, because portfolio *D* must lie on the plane described earlier, we could have obtained its expected return from the equation of the plane:

$$\bar{R}_i = 7.75 + 5(0.6) + 3.75(0.6) = 13$$

By the law of one price, two portfolios that have the same risk cannot sell at a different expected return. In this situation it would pay arbitrageurs to step in and buy portfolio *E* while selling an equal amount of portfolio *D* short. Buying portfolio *E* and financing it by selling *D* short would guarantee a riskless profit with no investment and no risk. We can see this quite easily. Assume the investor sells \$100 worth of portfolio *D* short and buys \$100 worth of portfolio *E*. The results are shown in the following table.

	Initial Cash Flow	End of Period Cash Flow	b_{i1}	b_{i2}
Portfolio *D*	+\$100	−\$113.0	−0.6	−0.6
Portfolio *E*	−\$100	\$115.0	0.6	0.6
Arbitrage portfolio	0	2.0	0	0

The arbitrage portfolio involves zero investment, has no systematic risk (b_{i1} and b_{i2}), and earns \$2. Arbitrage would continue until portfolio *E* lies on the same plane as portfolios *A*, *B*, and *C*.

We have established that all investments and portfolios must be on a plane in expected return, b_{i1}, b_{i2} space. If an investment were to lie above or below the plane, an opportunity would exist for riskless arbitrage. The arbitrage would continue until all investments converged to a plane.

The general equation of a plane in expected return, b_{i1}, b_{i2} space is

$$\bar{R}_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2} \tag{16.3}$$

This is the equilibrium model produced by the APT when returns are generated by a two-index model. Notice that λ_1 is the increase in expected return for a one-unit increase in b_{i1}. Thus λ_1 and λ_2 are returns for bearing the risks associated with I_1 and I_2, respectively.

More insight can be gained into the meaning of the λ_is by using Equation (16.3) to examine a particular set of portfolios. Examine a portfolio with b_{i1} and b_{i2} both equal to zero. The expected return on this portfolio equals λ_0. This is a zero-b_{ij} portfolio, and we denote its return by R_F. If the riskless asset is not available, R_F is replaced with $\bar{R}_Z$, the return on a zero-beta portfolio. Most researchers in this area assume that the intercept is in fact R_F.

Substituting $\bar{R}_F$ for λ_0 and examining a portfolio with a b_{i2} of 0 and a b_{i1} of 1, we see that

$$\lambda_1 = \bar{R}_1 - R_F$$

where $\bar{R}_1$ is the return on a portfolio having a b_{i1} of 1 and a b_{i2} of 0. In general, $\lambda_i = \bar{R}_j - R_F$ or λ_j is the expected excess return on a portfolio only subject to risk of index *j* and having a unit measure of this risk.

The analysis in this section can be generalized to the *J* index case:

$$R_i = a_i + b_{i1} I_1 + b_{i2} I_2 + \cdots + b_{iJ} I_J + e_i$$

By analogous arguments it can be shown that all securities and portfolios have expected returns described by the *J*-dimensional hyperplane

$$\bar{R}_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2} + \cdots + \lambda_J b_{iJ} \tag{16.4}$$

with $\lambda_0 = R_F$ and $\lambda_j = \bar{R}_j - R_F$.

A More Rigorous Proof of APT

Once again, we will derive APT assuming a two-index return-generating process. This derivation is sufficiently rich to allow generalization to any arbitrary number of indexes. The two-index model we use is that presented in Equation (16.2).

Taking the expected value of Equation (16.2) and subtracting it from Equation (16.2), we have

$$R_i = \bar{R}_i + b_{i1}\left(I_1 - \bar{I}_1\right) + b_{i2}\left(I_2 - \bar{I}_2\right) + e_i \tag{16.5}$$

Now a sufficient condition for an APT proof to hold is that there are enough securities in the market so that a portfolio with the following characteristics can be formed:

$$\sum_{i=1}^{N} X_i = 0$$

$$\sum_{i=1}^{N} X_i b_{i1} = 0$$

$$\sum_{i=1}^{N} X_i b_{i2} = 0$$

$$\sum_{i=1}^{N} X_i e_i \approx 0$$

The last condition is a requirement that residual risk be approximately zero.[5] The first of these four equations states that this portfolio involves zero investment. The remaining equations imply that this portfolio has no risk. This portfolio involves no investment and no risk; therefore it must produce an expected return of zero. In other words, the three equations plus the condition on residual risk just discussed imply that

$$\sum_{i=1}^{N} X_i \bar{R}_i = 0$$

Now there is another more mathematical interpretation of these equations. The equation

$$\sum_{i=1}^{N} X_i b_{i1} = 0$$

means that the vector of security proportions is orthogonal to the vector of b_{i1}s. Similarly, the first equation

$$\sum_{i=1}^{N} X_i = 0$$

means that the vector of security proportions is orthogonal to a vector of ones. We have just shown, in the previous paragraph, that if the vector of portfolio proportions is orthogonal to

[5]The assumption of zero residual risk might seem bothersome. Original proofs of APT assumed an infinite number of securities and well-diversified arbitrage portfolios. Because with uncorrelated residuals, each residual variance enters with a weight equal to the square of the fraction of money placed in that security, for well-diversified portfolios selected from an infinite or, in fact, a very large population of securities, residual risk will be very close to zero. A series of papers by Dybvig (1983), Grinblatt and Titman (1983, 1985), and Ingersoll (1984) investigate how closely the APT holds for finite economies and economies where residual risks are not uncorrelated. APT continues to hold, although it does not necessarily hold exactly the same for all securities (there can be very small errors for many securities, and there can be large pricing errors for a few securities).

a vector of ones, a vector of b_{i1}s, and a vector of b_{i2}s, this implies that the vector of security proportions is orthogonal to the vector of expected returns. But there is a well-known theorem in linear algebra that states that if the fact that a vector is orthogonal to $N - 1$ vectors implies it is orthogonal to the Nth vector, then the Nth vector can be expressed as a linear combination of the $N - 1$ vectors. In this case, the vector of expected returns can be expressed as a linear combination of a vector of ones, a vector of b_{i1}s, and a vector of b_{i2}s. Thus we can write the expected value for any security as a constant times 1, plus a second constant times b_{i1}, plus a third constant times b_{i2}, or

$$\overline{R}_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2}$$

This equation must hold for all securities and all portfolios. The λs can be evaluated by following the procedure used in the previous section of this chapter; namely, forming three portfolios with the characteristics

1. $b_{p1} = 0$ and $b_{p2} = 0$
2. $b_{p1} = 1$ and $b_{p2} = 0$
3. $b_{p1} = 0$ and $b_{p2} = 1$

we find that

$$\overline{R}_i = R_F + b_{i1}\left(\overline{R}_1 - R_F\right) + b_{i2}\left(\overline{R}_2 - R_F\right)$$

or for the general case,

$$\overline{R}_i = R_F + b_{i1}\left(\overline{R}_1 - R_F\right) + \cdots + b_{iJ}\left(\overline{R}_J - R_F\right)$$

Defining λ_0 as R_F and λ_j as $\overline{R}_j - R_F$, we can write this equation as

$$\overline{R}_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2} + \cdots + \lambda_J b_{iJ}$$

The principal strength of the APT approach is that it is based on the no-arbitrage condition. Because the no-arbitrage conditions should hold for any subset of securities, it is not necessary to identify all risky assets or a "market portfolio" to test the APT. It is reasonable to test it over a class of assets such as common stocks or even a smaller set such as the stocks making up the Standard and Poor's (S&P) index or all stocks on the New York Stock Exchange. One has to be somewhat careful in that the correct APT model for a larger class of securities can be different from (contain more influences than) an APT model appropriate for a smaller set of securities. Failure to find a model for a small set (type) of securities does not mean that a model does not exist across different types of securities. However, it is appropriate to use the APT to describe relative prices for a set of securities of interest to the investigator rather than deal with the whole population of risky assets. In fact, it has been argued that many tests of the CAPM were really tests of a single- or multiple-factor APT model.

An important characteristic of the APT theory is that it is extremely general. This generality is both a strength and a weakness. Although it allows us to describe equilibrium in terms of any multi-index model, it gives us no evidence as to what might be an appropriate multi-index model. Furthermore, APT tells us nothing about the size or the signs of the λ_js. This makes interpretation of tests difficult. We'll have more to say about this shortly.

ESTIMATING AND TESTING APT

The proof of any economic theory is how well it describes reality. Tests of APT are particularly difficult to formulate because all the theory specifies is a structure for asset

pricing: the economic or firm characteristics that should affect expected return are not specified. Let us review the structure of APT that will enter any test procedure.

We can write the multifactor return-generating process as

$$R_i = a_i + \sum_{j=1}^{J} b_{ij} I_j + e_i \tag{16.6}$$

The APT model that arises from this return-generating process can be written as

$$\overline{R}_i = R_F + \sum_{j=1}^{J} b_{ij} \lambda_j \tag{16.7}$$

It is worth spending a little time discussing the meaning of the variables b_{ij}, I_j, and λ_j.

Notice from Equation (16.6) that each security i has a unique sensitivity to each I_j but that any I_j has a value that is the same for all securities. Any I_j affects more than one security (if it did not, it would have been compounded in the residual term e_i). These I_js have generally been given the name *factors* in the APT literature. They are identical to the influences we called *indexes* in earlier chapters. The factors affect the returns on more than one security and are the sources of covariance between securities. The b_{ij}s are unique to each security and represent an attribute of the security. This attribute may be simply the sensitivity of the security to a particular factor, or it can be a characteristic of the security such as dividend yield.

Finally, from Equation (16.7), we see that λ_j is the extra expected return required because of a security's sensitivity to the jth attribute of the security. At this point the reader might note that Equation (16.6) looks suspiciously like the type of relationship we used in first-pass regression tests of the CAPM in Chapter 15, whereas Equation (16.7) bears a close resemblance to the type of equation used in second-pass tests. This intuition is correct. The problem is that, whereas for the CAPM, the correct I_j is defined (e.g., the excess return on the market portfolio for the simple CAPM), for the multifactor model and the APT, the set of I_js is not defined by the theory. To test the APT, one must test Equation (16.7), which means that one must have estimates of the b_{ij}s. Most tests of APT use Equation (16.6) to estimate the b_{ij}s. However, to estimate the b_{ij}s, we must have definitions of the relevant I_js. The most general approach to this problem is to estimate simultaneously factors (I_js) and firm attributes (b_{ij}s) for Equation (16.7). Most of the early tests of the APT employed this methodology. It still continues to be widely used in the finance literature and in practice. We examine this type of simultaneous estimation technique shortly. Before we do so, however, let us point out two alternative methods.

One alternative method is to specify a set of attributes (firm characteristics) that might affect expected return. When using this method, the b_{ij}s are directly specified. The b_{ij}s might include such characteristics as dividend yield and the firm's beta with the market. Once the b_{ij}s are specified, Equation (16.7) is used to estimate the λs and thus the APT model.

The second alternative method is to specify the factors I_js in Equation (16.6) and then to estimate the security attributes b_{ij}s and market prices of risk λ_js. Two approaches have been used to specify the factors. One approach is first to hypothesize (we hope on the basis of economic theory) a set of macroeconomic influences that might affect return and then to use Equation (16.6) to estimate the b_{ij}s. These influences might include variables such as the rate of inflation and the rate of interest.[6]

[6]BIRR has offered a commercial version of this research. A detailed description of their model can be found in Burmeister, Roll, and Ross (1994).

A second approach is to specify a set of portfolios as factors that the researcher believes captures the relevant influences affecting security returns. As in the previous case, Equation (16.6) is used to estimate the b_{ij}s, with the return on the hypothesized portfolios used as the I_js and b_{ij}s estimated via regression analysis. For either approach, Equation (16.7) is then estimated to obtain the λ_js and the associated APT model.

If any method other than factor analysis is used to obtain the b_{ij}s for testing APT, one is really conducting a joint test of the APT and the relevancy of the factors or characteristics that have been hypothesized as determining equilibrium. Each of these general approaches is now discussed in more detail.

Simultaneous Determination of Factors and Characteristics

A complete specification of Equation (16.6) would call for all factors (I_j) and attributes (b_{ij}) to be defined, so that the covariance between any residual return (the e_is not explained by the equation) was zero. Although it is not possible to produce this exact result, there is a body of statistical methodology that is very well suited to approximating this result. These techniques are called *factor analysis.* We present a simple example of a factor analytic solution in Appendix A to provide the reader who has not worked with this technique some feel for what it accomplishes.

Factor analysis determines a specific set of I_js and b_{ij}s such that the covariance of residual returns (returns after the influence of these indexes has been removed) is as small as possible.[7] In the terminology of factor analysis, the I_js are called *factors* and the b_{ij}s are called *factor loadings*. A specific factor analysis is performed for a specific number of hypothesized factors. By repeating this process for alternative hypotheses about the number of factors, a solution for two factors, three factors, . . . , and j factors is obtained. One can stop when the probability that the next factor explains a statistically significant portion of the covariance matrix drops below some level—for example, 50%.[8] Using this technique, it is not possible to be sure that one has captured all relevant factors. At best, statements such as the following can be made: "There is less than a 50% probability that another factor is needed." Whether one chooses to stop extracting factors when there is a 50% chance that no more are needed, or a 10% chance, or some other level is a matter of taste rather than mathematical rigor. Without a theory of how many factors should be present, the decision as to how many to extract from the data has to be made subjectively.

Factor analysis produces estimates of the factor loadings (b_{ij}) and the factors (I_j). Recall that the factor loadings b_{ij} are sensitivity measures and are like the β_is of the simple CAPM. At this point, a set of tests analogous to the first-pass regression tests discussed in Chapter 15 has been performed. The major difference is that one not only has identified the b_{ij}s but also has estimated how many factors (indexes) there should be and has determined the definition of each I_j. Each I_j is an index consisting of a (different) weighted average of the securities on which the factor analysis is performed.

---

[7]Principal component analysis is somewhat analogous to factor analysis. Recall from Chapter 8 that principal component analysis extracts from the data a set of indexes that best explains the variance of the data. Indexes are extracted in order of importance, and as many indexes are extracted as the smaller of the number of stocks or the number of observations. Factor analysis is covariance rather than variance driven. For a specified number of indexes it finds the set of that many indexes that best explains the covariance in the original data. There are alternative ways of performing factor analysis. Most empirical work in this area uses maximum likelihood factor analysis, and the techniques developed by Joreskog (1963, 1967, 1977) are often used.

[8]See Lawley and Maxwell (1963) for a discussion of the test procedure described. The reader should be aware that these tests are based on the assumption of multivariate normality. This is the procedure applied by Roll and Ross (1980).

The next step in testing the APT is to form a set of tests directly analogous to the second-pass tests performed by Fama and MacBeth (1973) on the simple CAPM.[9] By running a cross-sectional test, estimates of λs can be computed for each time period, and the average value of each λ_j and its variance over time can be computed. Roll and Ross (1980) were the first to perform this type of test. The mathematics of factor analysis allows this to be done more easily than with regression techniques, but the results are analogous to those that would be obtained by using the generalized least squares regression procedure. However, there are some problems with the use of factor analysis of which the reader should be aware. First, we have the same error-in-variables problem that we had when testing the standard CAPM. The factor loading b_{ij}s, like the betas from the first-pass regression, are estimated with error. This means that significance tests of λ_js are only asymptotically correct. There are three additional problems that are unique to factor analysis. First, there is no meaning to the signs of the factors produced by factor analysis, so the signs on the b_{ij}s and on the λ_js could be reversed. Second, the scaling of the b_{ij}s and the λ_js is arbitrary. For example, all b_{ij}s could be doubled and the resultant λ_js halved. Third, there is no guarantee that factors are produced in a particular order, so when analysis is performed on separate samples, the first factor from one sample may be the third from another sample.

The procedure discussed is that used by Roll and Ross (1980) in their classic study of APT. They applied factor analysis to 42 groups of 30 stocks using daily data for the time period July 3, 1962 to December 1972. The results of their first-pass test are rather striking. These tests show that, in over 38% of the groups, there was less than a 10% chance that a sixth factor had explanatory power, and in over three-fourths of the groups, there was a 50% chance that five factors were sufficient. While Roll and Ross try several different second-pass tests, their major results are that at least three factors are significant in explaining equilibrium prices but that it is unlikely that four are significant. On the surface it would appear that they find more factors significant than one would expect to find under the standard CAPM model or the zero-beta version of the CAPM.

In Japan, APT has been tested and shows a clear superiority over the CAPM in selecting securities as well as in explaining past returns. For example, Elton and Gruber (1982, 1988) find that a five-factor APT model does a better job of explaining and predicting expected returns than does a single-factor or CAPM model. In particular, in the Japanese stock market the CAPM model appears to break down. In Japan, unlike other markets, small stocks have smaller betas than large stocks. This should imply a lower expected return given the CAPM, and yet small stocks have significantly higher excess returns. This happens when *small* is defined as anything but the largest 100 stocks on the Tokyo Stock Exchange. These problems are not nearly as great when a multifactor model is used. Furthermore, a multifactor model does a much better job of allowing mimicking portfolios to be constructed (as both index funds and hedge portfolios for futures and option trading) than does a single-index model. The APT model is almost universally used by industry as a replacement for the CAPM model in Japan.

An Alternative Approach to Testing the APT

If we could specify a priori either the factors that affected stock returns or the characteristics of stocks that affected returns, we would then have a much easier estimation problem to solve. A debate exists among academics and practitioners about whether part of

[9]Alternate tests such as those advocated by Gibbons (1981), described in the previous chapter, or those advocated by Burmeister et al. (1988), described later in this chapter, can be used instead of the second-pass test.

the model should be prespecified on the basis of theory or whether all of the parameters should be determined empirically. This type of debate has gone on since the dawn of modern science. The issue is discussed by Roll and Ross (1980). They state that "we do consider the basic underlying causes of the generating process of returns to be a potentially important area of research but we think it is an area that can be investigated separately from testing asset pricing theories." The problem is that, without a theory, the empirical tools one uses are a lot weaker and the results of tests harder to interpret. For example, in the APT, we have no idea of what the size or even the sign of factor prices should be. All we can say is that we expect some of them to be statistically different from zero. On the other hand, in the Sharpe–Lintner CAPM, the price of beta was supposed to be $\bar{R}_m - R_F$, a quantity that we expected to be positive and about which we have some rough idea of magnitude.

The controversy we are discussing would be easy to resolve if we had a theory of the appropriate factors or characteristics that determine security returns. Someday we hope to have one. In the absence of such a theory, all we can do is examine three attempts to prespecify one set of variables in the multifactor model. One attempt hypothesized a set of firm characteristics, another hypothesized a set of macroeconomic indexes, and the third specifies a set of portfolios as the indexes.

Specifying Attributes of Securities

In the preceding section of this chapter we examined the use of maximum likelihood factor analysis to determine simultaneously the characteristics that affect return and the extra return required because of a security's sensitivity to these characteristics. If a set of characteristics that affects return could be specified a priori, then the market price of these characteristics over any period of time could be measured fairly easily.

The estimating equation would be of the form

$$\bar{R}_i = \lambda_0 + \lambda_1 b_{i1} + \lambda_2 b_{i2} + \cdots + \lambda_J b_{iJ}$$

for the case of J characteristics. In this equation the b_{ij}s would be the value each characteristic took on, and the λ_js the average extra return required because of these characteristics. The values of the λ_js would be estimated via regression analysis. This procedure is directly analogous to a second-pass test of the CAPM.

One model using multiple firm characteristics has been constructed and tested by Sharpe (1982). He starts with the hypothesis that equilibrium returns should be affected by the following characteristics: a stock's beta with the S&P index, its dividend yield, the size of the firm (market value of equity), its beta with long-term bonds, its past value of alpha (the intercept of the regression of past excess returns against excess returns on the S&P index), and eight-sector membership variables. Sharpe does not attempt an elaborate economic rationale for these variables but rather states that he has selected them more or less "ex cathedra." We would expect both beta and dividend yield to be related positively to expected returns based on the theory discussed in Chapters 13, 14, and 15. Size may well be, at least in part, a proxy for liquidity. If so, size should enter the model with a negative sign. If sensitivity to interest rates is an important variable, we would expect bond beta to play a role in determining equilibrium returns. If the past value of alpha proves significant, it would be evidence of autocorrelation of the residuals from the CAPM. This might indicate that there are some added variables explaining cross-sectional returns that were not captured in the model. The use of sector membership as an additional set of variables implies that membership in a particular sector of the economy has an important effect on equilibrium return.

Table 16.1 Cross-Sectional Data on Sharpe's Multifactor Model

Attribute	Annualized Value of Associated λ	Percentage of Months in Which Associated λ Was Significantly Different from Zero
Beta	5.36	58.3
Yield	0.24	39.5
Size	−5.56	56.5
Bond beta	−0.12	28.2
Alpha	−2.00	43.5
Sector Membership		
Basic industries	1.65	32.5
Capital goods	0.16	18.7
Construction	−1.59	15.3
Consumer goods	−0.18	39.3
Energy	6.28	36.9
Finance	−1.48	16.3
Transportation	−0.57	43.9
Utilities	−2.62	35.0

The results of applying this model to 2,197 stocks on a monthly basis for all months between 1931 and 1979 are summarized in Table 16.1, which reports the average coefficients (on an annualized basis) over the entire period and the percentage of months in which the coefficients were significantly different from zero at the 5% level. Note that for those variables where we had clear expectations about the sign of the relationship and return, our expectations are borne out. Furthermore, note that although on the basis of chance, we would expect any firm characteristic to be significant about 5% of the time, each characteristic was significant a much higher percentage of the time.

Sharpe seems to have identified some additional characteristics, beyond a stock's beta with a proxy for the market portfolio, that are useful for explaining cross-sectional returns over time. He recognizes that his model is rather ad hoc in nature, but it is an indication that increased research into significant economic characteristics of a stock might allow us to build better models of equilibrium.

Specifying the Influences Affecting the Return-Generating Process

Another alternative to the joint determination of factor loadings and factors discussed in the earlier section of this chapter is the specification (one hopes on the basis of economic theory) of the set of influences or indexes (I_js) that should enter the return-generating process.

Chen, Roll, and Ross (1986) have hypothesized and tested a set of economic variables. They reason that return on stocks should be affected by any influence that affects either future cash flows from holding a security or the value of these cash flows to the investor (e.g., changes in the appropriate discount rate on future cash flows). Chen, Roll, and Ross construct sets of alternative measures of unanticipated changes in the following influences:

1. *Inflation.* Inflation impacts both the level of the discount rate and the size of the future cash flows.

2. *The term structure of interest rates.* Differences between the rate on bonds with a long maturity and a short maturity affect the value of payments far in the future relative to near-term payments.
3. *Risk premia.* Differences between the return on safe bonds (AAA) and more risky bonds (BAA) are used to measure the market's reaction to risk.
4. *Industrial production.* Changes in industrial production affect the opportunities facing investors and the real value of cash flows.

Chen, Roll, and Ross then examined these measures or indexes

1. to see if they were correlated with the set of indexes extracted by the factor analysis used by Roll and Ross as described in a previous section of this chapter
2. to see if they explained equilibrium returns

When they examine the relationship between the macroeconomic variables and the factors (indexes) over the period to which the factors were formed (fit), they find a strong relationship. Furthermore, when the relationship is tested over a holdout period (a period following the fit period), the relationship continues to be strong. There appears to be a significant relationship between the hypothesized macroeconomic variables and the statistically identified systematic factors in stock market returns.

The second set of tests involves investigating whether returns are related to the sensitivity of a stock to their macroeconomic variables. The procedure is analogous to the two-step procedure used by Fama and MacBeth (and discussed in the previous chapter) to investigate the CAPM. In the first stage, time series regressions are run for each of a series of portfolios to estimate each portfolio's sensitivity to each macroeconomic variable [the b_{ij}s of Equation (16.6)]. Then the market price of risk [the λ_js of Equation (16.7)] is estimated by running a cross-sectional regression each month and looking at the average of the market price in each month. Chen, Roll, and Ross find that the macrovariables are significant explanatory influences on pricing. Furthermore, when the beta of each portfolio with the market was introduced as an additional variable along with the sensitivity of each portfolio to the macroeconomic variables, it did not show up as significant in the second-stage (cross-sectional) regression.

Chen, Roll, and Ross recognize that they cannot claim to have found the (correct) state variables for asset pricing. However, they certainly have made an important start in that direction.

Their work is continued in a series of papers by Burmeister and McElroy. Burmeister and McElroy have integrated tests of the factor models, CAPM, and APT. It is worthwhile reviewing two of their tests. The first test is constructed using the multi-index model described in Chapter 8. More specifically, returns are assumed to be generated by the following five indexes (see Chapter 8):

I_1 = default risk as measured by the return on long-term government bonds minus the return on long-term corporate bonds plus one-half of 1%

I_2 = time premium as measured by the return on long-term government bonds minus the one-month Treasury bill rate one month ahead

I_3 = deflation as measured by expected inflation at the beginning of the month minus actual inflation during the month

I_4 = change in expected sales

I_5 = the market return not captured by the first four variables

The fifth variable is a proxy for any unobserved general influences. As explained in Appendix B, it is estimated by taking the residuals from a regression of a diversified portfolio

(the authors use the S&P composite index) against the first four observable variables described earlier. The regression the authors found was

$$R_M - \lambda_0 = 0.00224 - 1.330I_1 + 0.558I_2 + 2.286I_3 - 0.935I_4$$
$$(0.619) \quad (-3.94) \quad (4.96) \quad (1.997) \quad (-2.27)$$

The first four factors account for about 25% ($R^2 = 0.24$) of the variation in the return on the S&P composite index, and each of the four coefficients is significant.

When the sensitivities (b_{ij}) are estimated for each firm, more than two-thirds of the sensitivities are statistically different from zero at the 5% level, and the five variables typically account for 30% to 50% of the variation of returns of individual firms. In general, b_{i1} appears with a significant negative coefficient, whereas b_{i2} and b_{i5} appear with significant positive coefficients. The remaining two variables have a more ambiguous impact on stock returns.

The prices (λ_i) of each of the five sensitivities implied by the model are all positive and all statistically significantly different from zero. The average value of the λs using monthly returns is contained in the following table:

	Mean λ Value	*t* Statistic
λ_1	0.44	4.27
λ_2	1.00	4.76
λ_3	0.04	1.83
λ_4	0.15	2.21
λ_5	0.51	3.21

Burmeister and McElroy go on to test whether the model they find is a return-generating model or an APT model. If it is an APT model, then the intercept should be a constant which implies no arbitrage. When they impose this constraint, they find no significant decrease in explanatory power, which is evidence in support of an APT model.

Specifying a Set of Portfolios Affecting the Return-Generating Process

Another alternative is to specify a set of portfolios (I_js) (which may or may not include the market portfolio) that a priori are thought to capture the influences affecting security returns. These portfolios are selected on the basis of a belief about the types of securities and/or economic influences that affect security returns.[10]

An example of this type of approach is that used by Fama and French (1993) to construct a model to explain returns and expected returns on both stocks and bonds. In addition to using the returns on a market portfolio of stocks, they use the returns on other portfolios to represent the I_js in the return-generating process. These portfolios are

[10]We should point out that this is fundamentally different from the approach of factor-replicating portfolios that has been discussed by Lehmann and Modest (1988) and Huberman, Kandel, and Stambaugh (1987), among others. In these approaches, either factor analysis is used to extract factors or macroeconomic variables are hypothesized as important, and then a mathematical programming problem is solved to find portfolios that mimic the underlying factors.

1. the difference in return on a portfolio of small stocks and a portfolio of large stocks (small minus large)
2. the difference in return between a portfolio of high-book-to-market stocks and a portfolio of low-book-to-market stocks (high minus low)
3. the difference between the monthly long-term government bond return and the one-month Treasury bill return
4. the difference in the monthly return on a portfolio of long-term corporate bonds and a portfolio of long-term government bonds

Note that all variables are either the return on portfolios of assets or the difference in the return of two portfolios of assets.[11] The latter can be considered a portfolio with a set of stocks sold short. Clearly this model has elements in common with the models that have been presented earlier in this chapter. We saw that Chen, Roll, and Ross and Burmeister and McElroy use bond return variables similar to those used in this model. Whether one describes these as measures of macroeconomic variables or portfolios is largely a matter of taste. The unique aspect of this model is in the formulation of the variables representing size and book-to-market ratios. In Sharpe's model (described earlier), size enters as a firm characteristic or a b_{ij}. Size is measured in dollars (actually the natural logarithm of dollars), and a λ is associated with it via cross-sectional regression. What Fama and French have done is to convert the size component from a direct measure to a return concept by constructing a portfolio to capture this influence. The b_{ij} associated with size is not the log of size for any company i but rather is the sensitivity of that company to the return on the size portfolio. Because size is measured by the return on a portfolio, it now enters the return-generating process as well as the pricing equation. This allows Fama and French to investigate both the time series and cross-sectional properties of size.

Fama and French test the model described previously in a number of time series tests. The cross-sectional implications are tested by examining whether the intercepts of the time series of excess returns indeed equal zero, as APT would suggest. They find that, in fact, the intercepts are zero and that this portfolio model is successful in explaining expected stock returns. More specifically, they conclude that at a minimum, our results show that five factors do a good job explaining a) common variations in bond and stock returns and b) the cross-section of average returns.

At this point, the Fama and French approach has become the standard multifactor model used extensively both by academics and practitioners. Ready access to these factors through Kenneth French's data library[12] and the large number of research papers that have been written using these factors has made this the model of choice in current empirical research in finance. Following Fama and French (1992), in equity market research, three factors—the market return in excess of short-term Treasury bill return, the return on small equity minus large equity (SMB), and the return on high-market-to-book less low-market-to-book stocks (HML)—are most commonly used. Following Jegadeesh and Titman (1993, 2003), this set of three factors is sometimes augmented by a momentum factor defined analogously using the difference between the return on stocks that performed well over the previous 12 months and the return on stocks that

[11]Elton, Gruber, Das, and Hlavka (1993), Blake, Elton, and Gruber (1993), and Elton, Gruber, and Blake (1994) investigate alternative return-generating processes. In the latter paper, they develop an APT model where some or all of the indexes represent portfolios of assets. This work is discussed more fully in Chapter 25.

[12]http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

performed poorly over the same period (UMD).[13] Part of the popularity of these factors can be explained by the association of book to market as an index of value among both academics and practitioners.[14] Vuolteenaho (2002) finds that almost all of the time series variation in the aggregate book to market ratio is to be explained in terms of changes in expected returns and argues that recent returns, return on equity, and book to market jointly predict returns. It is unclear why these measures of value are not priced into the value of the equities. The main reason, however, for the widespread acceptance of these factors is the empirical result documented by Fama and French (1992) and others that exposure to these factors explains a significant fraction of the cross-sectional dispersion in returns.

However, the book to market factor is not the only factor that can explain future returns. As Fama and French (2006) observe, valuation theory says that expected stock returns are related to three variables: book to market, expected profitability, and expected investment. Indeed, in a recent paper, Nagel (2012) argues that the choice of market factor, SMB, and HML as factors is somewhat ad hoc. Kogan and Tian (2012) argue that it is possible to define as many as 351 possible three-factor linear pricing models that match return spreads associated with as many as 15 out of 27 commonly used firm characteristics. This concern is compounded by the fact that SMB and HML are constructed from portfolios that span the very same expected return spreads along the size and book to market dimensions that the model is trying to explain. This would amount to a tautology were it not for the fact that SMB and HML explain not only cross-sectional dispersion of returns but also their time series characteristics. While Fama and French (1993) conjectured that their factors could proxy for macroeconomic factors and that these factor exposures may explain otherwise anomalous return patterns, Nagel (2012) argues that the empirical evidence that has accumulated since Fama and French (1996) suggests that there are other important sources of cross-sectional variation in returns unrelated to these factors.

An alternative approach is to consider the macroeconomic models more directly. As described in Chapter 15, consideration of the multiperiod consumption-investment model gives rise to a single-factor consumption asset pricing model where expected returns are linearly related to the logarithm of consumption growth. The empirical evidence in favor of this consumption-based model has been disappointing. Lettau and Ludvigson (2001) argue that the reason for this poor performance is a result of the fact that this asset pricing model is strictly conditional in nature. Using the log consumption-wealth ratio (*cay*) as a conditioning variable allows them to consider the single-factor conditional model in terms of a multifactor unconditional model. This extension leads to a dramatic improvement in the explanatory power of the model. However, the log consumption-wealth variable is not the only possible conditional variable. Lustig and van Nieuwerburgh (2005) consider the housing collateral ratio (*my*) as a potentially conditioning variable and derive a similar three-factor unconditional model. Santos and Veronesi (2006) consider the labor income to consumption ratio (s^w); Li, Vassalou, and Xing use investment growth rates for households (ΔI_{HH}), nonfinancial corporations

[13]This is particularly the case in managed fund research, following the work of Carhart (1997). Ang, Chen, and Xing (2001) argue that this momentum effect may represent a reward for bearing downside risk. The UMD factor is also available through the Kenneth French data library.

[14]See, for example, Rosenberg, Reid, and Lanstein (1985).

(ΔI_{corp}), and the noncorporate sector (ΔI_{NCorp}); and Yogo (2006) uses the growth rates in durable (ΔC_{Dur}) and nondurable (ΔC_{Ndur}) consumption. These are just some examples of the macroeconomic factors that arise naturally out of the multiperiod consumption and investment model.

One of the limitations of these multifactor macroeconomic factor–based models is that the data to estimate them are available on at most a quarterly basis. Nevertheless, it is possible to compare them directly with each other and with the Fama and French and standard CAPM models. Results using quarterly data from 1963 to 2004 are reported in Lewellen, Nagel, and Shanken (2010) and are reproduced in Table 16.2. Considering first the results based on 25 Fama and French (1992) portfolios, the Fama and French model compares favorably with the macroeconomic-based multifactor models, while the standard CAPM and consumption-based asset pricing models perform relatively poorly. Nagel (2012) argues that this result is not altogether surprising as the SMB and HML factors are constructed using the same portfolio returns the model is meant to explain. Furthermore, one interpretation of the large and statistically significant value of the Gibbons, Ross, and Shanken T^2 statistic is that the Fama and French model is misspecified as a description of equibrium pricing. Another interpretation is that these results simply confirm results found in earlier tests of the CAPM that the intercept in excess return regressions appears to be positive. Frazzini and Pedersen (2011) argue that we should expect this result given significant margin and leverage constraints in U.S. equity markets. Expanding the set of portfolios to consider 20 industry-based portfolios in addition to the 25 Fama and French (1992) portfolios leads to a substantial fall in the explanatory power of all models considered, and the low value of the generalized least squares R^2 suggests that none of these models has a great deal of explanatory power once cross-sectional covariances are accounted for. Note, however, that the confidence intervals for these statistics are rather large, consistent with the fact that these tests are based on quarterly data over a period of substantial changes in the U.S. financial system.

The table reports slopes, Shanken (1992) *t*-statistics (in parentheses), and other statistics from cross-sectional regressions of average excess returns on estimated factor loadings for eight asset pricing models. Returns are quarterly, in percent. The test assets are Fama and French's 25 size-B/M portfolios used alone (FF25) or together with their 30 industry portfolios (FF25+30 ind.). The OLS R^2 is an adjusted R^2. The cross-sectional T^2 statistic tests whether pricing errors in the cross-sectional regression are all zero, with simulated *p*-values in brackets. Ninety-five percent confidence intervals for the true R^2s are reported in brackets next to the sample values. The models considered are those of Lettau and Ludvigson (2001) (LL), Lustig and van Nieuwerburgh (2005) (LVN), Santos and Veronesi (2006) (SV), Yogo (2006), the standard CAPM, the Consumption CAPM, and finally, the Fama and French (1992) (FF) models. These models are estimated from 1963 to 2004, except Yoga's (2006) model, which uses factor data through 2001. The variables are *cay*, Lettau and Ludvigson's (2001) consumption-to-wealth ratio; Δc, the log consumption growth; *my*, Lustig and Van Nieuwerburgh's (2004) housing-collateral ratio based on mortgage data; R_M, the CRSP value-weighted excess return; s^w, labor income to consumption ratio; ΔI_{HH}, ΔI_{corp}, ΔI_{Ncorp}, the log investment growth for households, nonfinancial corporations, and the noncorporate sector, respectively; ΔC_{Ndur}, ΔC_{Dur}, Yoga's (2006) log consumption growth for nondurables and durables, respectively; and finally SMB, HML, Fama and French's (1992) size and B/M factors. (Lewellen, Nagel, and Shanken 2010.)

Table 16.2 Empirical Tests of Multifactor Asset Pricing Models 1963–2004

Models	Variables								OLS R^2		GLS R^2		T^2	
LL(2001)	Const		*cay*		Δc		*cay**Δc							
FF25	3.33	(2.28)	0.81	(1.25)	0.25	(0.84)	0.00	(0.42)	0.58	[0.30,1.00]	0.05	[0.00,0.50]	33.9	[p=0.08]
FF25+30 ind.	2.50	(3.29)	0.48	(1.23)	0.09	(0.53)	0.00	(−1.10)	0.00	(0.00,0.35]	0.01	[0.00,0.20]	193.8	[p=0.00]
LVN(2005)	Const		*my*		Δc		*my**Δc							
FF25	3.58	(2.22)	4.23	(0.76)	0.02	(0.04)	0.10	(1.57)	0.57	[0.35,1.00]	0.02	[0.00,0.35]	20.8	[p=0.57]
FF25+30 ind.	2.78	(3.51)	0.37	(0.13)	0.02	(0.09)	0.03	(1.40)	0.09	[0.00,1.00]	0.00	–	157.1	[0=0.04]
SV(2006)	Const		R_M		R_M*s^w									
FF25	3.58	(1.96)	−0.95	(−0.58)	−0.21	(−2.06)			0.27	[0.00,1.00]	0.02	[0.00,0.40]	26.0	[p=0.63]
FF25+30 ind.	2.57	(2.77)	0.49	(0.44)	−0.09	(−1.99)			0.08	[0.00,1.00]	0.02	[0.00,0.40]	160.8	[p=0.07]
LVX(2006)	Const		ΔI_{HH}		ΔI_{Corp}		ΔI_{Ncorp}							
FF25	2.47	(2.13)	−0.80	(−0.39)	−2.65	(−1.03)	−8.58	(1.96)	0.80	[0.75,1.00]	0.26	[0.05,1.00]	25.2	[p=0·29]
FF25+30 ind.	2.22	(3.14)	0.20	(0.19)	−0.93	(−0.58)	−5.11	(2.32)	0.42	[0.20,1.00]	0.04	[0.00,0.55]	141.2	[p=0.11]
Yogo(2006)	Const		ΔN_{Ndur}		ΔC_{Dur}		RM							
FF25	1.98	(1.36)	0.28	(1.00)	0.67	(2.33)	0.48	(0.29)	0.18	[0.00,1.00]	0.04	[0.00,0.55]	24.9	[p=0.69]
FF25+30 ind.	1.95	(2.27)	0.18	(1.01)	0.19	(1.52)	0.12	(0.11)	0.02	[0.00,0.60]	0.05	[0.00,1.00]	159.3	[p=0.06]
CAPM	Const		R_M											
FF25	2.90	(3.18)	0.44	(0.39)					−0.03	[0.00,0.55]	0.01	[0.00,0.25]	77.5	[p=0.00]
FF25+30 ind.	2.03	(2.57)	0.10	(0.09)					−0.02	[0.00,0.35]	0.00	[0.00,0.05]	225.2	[p=0.00]
Cons.CAPM	Const		Δc											
FF25	1.70	(2.47)	0.24	(0.89)					0.05	[0.00,1.00]	0.01	[0.00,0.50]	60.6	[p=0.01]
FF25+30 ind.	2.07	(3.51)	0.03	(0.15)					−0.02	[0.00,0.65]	0.00	–	224.5	[p=0.00]
FF (1992)	Const		R_M		5MB		HML							
FF25	2.99	(2.33)	−1.42	(−0.98)	0.80	(1.70)	1.44	(3.11)	0.78	[0.60,1.00]	0.19	[0.05,0.65]	56.1	[p=0.00]
FF25+30 ind.	2.21	(2.14)	−0.49	(−0.41)	0.60	(1.24)	0.87	(1.80)	0.31	[0.00,0.90]	0.06	[0.05,0.35]	200.4	[p=0.00]

APT AND CAPM

Before continuing our examination of APT models, we should discuss the fact that the APT model—and, in fact, the existence of a multifactor model, including one where more than one factor is priced—is not necessarily inconsistent with the Sharpe–Lintner–Mossin form or one of the other forms of the CAPM.

The simplest case in which an APT model is consistent with the simple form of the CAPM is the case where the return-generating function is of the form

$$R_i = a_i + \beta_i R_m + e_i$$

If returns are generated by a single-index model, the single index is the return on the market portfolio, and a riskless rate exists, then the methodology at the beginning of the chapter can be used to show that

$$\overline{R}_i = R_F + \beta_i\left(\overline{R}_m - R_F\right)$$

If the return-generating function is more complex than this, does it imply that the simple CAPM cannot hold? The answer is no. Recall that the simple CAPM does not assume that the market is the only source of covariance between returns. Let us assume that the return-generating function is of the multi-index type:

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + e_i \tag{16.8}$$

The indexes can be industry indexes, sector indexes, or indexes of broad economic influences such as the rate of inflation. All we assume is that the set of indexes used captures all the sources of covariance between securities [e.g., $E(e_i e_j) = 0$].

The APT equilibrium model for this multifactor return-generating process with a riskless asset is

$$\overline{R}_i = R_F + b_{i1}\lambda_1 + b_{i2}\lambda_2 \tag{16.9}$$

Recall that if the CAPM is the equilibrium model, it holds for all securities as well as all portfolios of securities. Assume the indexes can be represented by portfolios of securities. Actually, we have seen that λ_j is the excess return on a portfolio with a b_{ij} of 1 on one index and a b_{ij} of zero on all other indexes. If the CAPM holds, the equilibrium return on each λ_j is given by the CAPM or

$$\lambda_1 = \beta_{\lambda 1}\left(\overline{R}_m - R_F\right)$$
$$\lambda_2 = \beta_{\lambda 2}\left(\overline{R}_m - R_F\right)$$

Substituting into Equation (16.9) yields

$$\overline{R}_i = R_F + b_{i1}\beta_{\lambda 1}\left(\overline{R}_m - R_F\right) + b_{i2}\beta_{\lambda 2}\left(\overline{R}_m - R_F\right)$$
$$\overline{R}_i = R_F + \left(b_{i1}\beta_{\lambda 1} + b_{i2}\beta_{\lambda 2}\right)\left(\overline{R}_m - R_F\right)$$

Defining β_i as $(b_{i1}\beta_{\lambda 1} + b_{i2}\beta_{\lambda 2})$ results in the expected return of $\overline{R}_i$ being priced by the CAPM.

$$\overline{R}_i = R_F + \beta_i\left(\overline{R}_m - R_F\right)$$

The APT solution with multiple factors appropriately priced is fully consistent with the Sharpe–Lintner–Mossin form of the CAPM.

We wish to stress this point. Employing the Roll and Ross procedure and finding that more than one λ_j is significantly different from zero is not sufficient proof to reject any

CAPM. If the λ_js are not significantly different from $\beta_{\lambda j}(\bar{R}_m - R_F)$, the empirical results could be fully consistent with the Sharpe–Lintner–Mossin form of the CAPM. It is perfectly possible that more than one index explains the covariance between security returns but that the CAPM holds.

Although we have demonstrated this with the simple CAPM, it should be apparent to the reader that other values of λ_js can exist that are fully consistent with the more complex nonstandard forms of the CAPM reviewed in Chapter 14.

RECAPITULATION

The APT theory remains the newest and most promising explanation of relative returns. The theory promises to supply us with a more complete description of returns than the CAPM. Recent work, some of which employs a set of macro variables and some of which employs a set of portfolios, is quite encouraging. The fact that a number of studies have found a set of macro variables and portfolios that impact average returns and are not only priced but also priced differently than the CAPM would imply is of both practical and theoretical significance. One word of caution is in order. It is possible that these additional influences are priced not because the APT is the correct model for expected returns but because we have not correctly identified the market in constructing our model. The residual market plus the other variable employed in the model may together simply serve as a proxy for the (true but unobserved) market in the manner suggested in Chapter 15. Even if this is correct, the use of these multi-index models is, on a practical level, a better explanation for returns than any of the market proxies that have been proposed to date.

A section on the uses of multi-index models and APT follows. Although there are many reasons for adding this section, most of which are discussed later, perhaps the key reason is that after we teach APT, so many of our students remark that it seems more complex than the CAPM and ask why we bothered with it.

Multi-index Models, APT, and Portfolio Management

The use of multi-index models and multi-index equilibrium models (APT models) in the selection of securities and the management and evaluation of portfolios is growing rapidly. Many brokerage firms, financial institutions, and financial consulting firms have developed their own multi-index models to aid in the investment process. These models have become increasingly popular because they allow risk to be more tightly controlled and they allow the investor to protect against specific types of risk to which she is particularly sensitive or to make specific bets on certain types of risk.

In this section we discuss the use of APT and multi-index models to aid in passive management, active management, and portfolio evaluation. Before we do so, we review multi-index models and APT briefly and present a simple example of an APT model that we use to illustrate some of the phenomena we discuss in this section.

Review of Multi-index Models and APT

Earlier in this chapter we presented a return-generating process that expressed the return on any security as a linear function of a series of indexes:

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + b_{i3}I_3 + \cdots + e_i \tag{16.1a}$$

It is convenient for purposes of this section to assume that each index has been either formulated or adjusted to have a mean equal to zero. Because the indexes and residuals have a mean of zero, taking the expected value of both sides of Equation (16.1) results in

$$\overline{R}_i = a_i$$

Thus setting the mean of each index equal to zero has the effect of ensuring that a_i is equal to the expected return on security *i*.

We saw that Equation (16.1a) leads to a description of expected returns given by

$$\overline{R}_i - R_F = \lambda_1 b_{i1} + \lambda_2 b_{i2} + \lambda_3 b_{i3} + \cdots \tag{16.10}$$

where b_{ij}s represent the sensitivity of a security's return to index *j* and are a measure of the risks inherent in the security under study and λs represent the reward for bearing these risks (price of risk).

Combining Equations (16.1) and (16.10) by recognizing that $a_i = \overline{R}_i$,

$$\begin{aligned} R_i &= R_F + \lambda_1 b_{i1} + \lambda_2 b_{i2} + \lambda_3 b_{i3} + \cdots \\ &+ b_{i1} I_1 + b_{i2} I_2 + b_{i3} I_3 + \cdots + e_i \end{aligned} \tag{16.11}$$

There are several ways of identifying the *I*s in Equation (16.1) and the b_{ij}s and λ_js in Equation (16.11). However, a specific model will help illustrate the use of these types of models.

Let us assume that we have identified four influences in the return-generating model (Equation 16.1) and that

I_1 = unexpected change in inflation, denoted by I_I

I_2 = unexpected change in aggregate sales, denoted by I_S

I_3 = unexpected change in oil prices, denoted by I_O

I_4 = the return in the S&P index constructed to be orthogonal to the other influences, denoted by I_M

Furthermore, assume that oil risk is not priced ($\lambda_O = 0$). Equation (16.10) becomes

$$\overline{R}_i - R_F = \lambda_I b_{iI} + \lambda_S b_{iS} + \lambda_M b_{iM}$$

whereas Equation (16.11) becomes

$$R_i - R_F = \lambda_I b_{iI} + \lambda_S b_{iS} + \lambda_M b_{iM} + b_{iI} I_I + b_{iS} I_S + b_{iO} I_O + b_{iM} I_M + e_i$$

Recall that all *I*s have an expected value of zero.[15]

The set of λs on these factors consistent with the results reported by Burmeister, Roll, and Ross are

$\lambda_I = -4.32$

$\lambda_S = 1.49$

$\lambda_M = 3.96$

[15]The model we describe here and the values we present represent a simplified version of the model and parameters described in Burmeister, Roll, and Ross (1994). Their model contains additional influences to those cited and does not contain an oil index. We wanted to include an unpriced index to show the role of unpriced indexes in portfolio management. The Salomon Brothers risk index model, discussed in Chapter 8, on multi-index models, also does not include an oil index for U.S. stocks, though they find this index is an important influence in Japan, the United Kingdom, Germany, and France.

while the sensitivities (b) values for the S&P index were

$$b_{S\&P\ I} = -0.37$$
$$b_{S\&P\ S} = 1.71$$
$$b_{S\&P\ O} = 0.00$$
$$b_{S\&P\ M} = 1.00$$

The parameterization of the model allows us to recognize the importance of any factor in determining the expected excess return on the S&P index. To do so, simply multiply the b associated with a factor times the associated price of risk (λ).

Factor	b	λ	Contribution to S&P Expected Excess Return (%)
Inflation	−0.37	−4.32	1.59
Sales growth	1.71	1.49	2.54
Oil prices	0.00	0.00	0.00
Market	1.00	3.96	3.96
Expected excess return for S&P index			8.09

This table shows that the expected excess return (return above the riskless rate) for the S&P index is 8.09%. Sales growth contributes 2.54% to the expected return for the S&P. In other words, sensitivity to sales growth accounts for 2.54 ÷ 8.09 or 31.4% of the total expected excess return.

The same type of analysis can be used to examine the importance of the sources of risk for the expected excess return on any security or portfolio. For example, for a portfolio of growth stocks, the bs, λs, and contribution to expected excess return are shown later:[16]

Factor	b	λ	Contribution to Growth Stock Portfolio Expected Excess Return (%)
Inflation	−.50	−4.32	2.16
Sales growth	2.75	1.49	4.10
Oil prices	−1.00	0.00	0.00
Market	1.30	3.96	5.15
Expected excess return for growth stock portfolio			11.41

Notice that the expected excess return for the growth stock portfolio (11.41) is higher than it was for the S&P index (8.09). This is not surprising because the growth stock portfolio has more risk, with respect to each index, than the S&P portfolio.[17]

Individual influences (indexes) have a different absolute and relative contribution to the expected excess return on a growth stock portfolio than they have on the S&P index. For example, the contribution of sales growth to expected excess return is now 4.10%. Sales

[16]Although estimating the cost of equity capital falls beyond the scope of this book, the preceding analysis leads naturally to estimates of cost of capital. For example, the cost of capital of any stock or portfolio can be found by adding the riskless rate to the estimate of excess return from the APT model. For growth stocks this would be $R_F + 11.41$. For a detailed explanation of using APT to determine cost of equity capital, see Elton, Gruber, and Mei (1994).

[17]Note that all b values, except for sensitivity to inflation, are larger for the growth stock portfolio than for the S&P portfolio. Though the b value for inflation is smaller for the growth stock portfolio, this portfolio is still less desirable with respect to inflation sensitivity because (unlike other λs) the price of inflation sensitivity (λ_1) is negative.

growth accounts for 35.9% of the excess return on the growth stock portfolio. It is not surprising that growth stocks are more sensitive to sales growth than the typical stock. What might be surprising, though it is generally true, is that growth stocks are more sensitive to all important indexes. So although the increase in sensitivity to sales growth causes the largest increase in expected excess returns, changes in all influences lead to greater excess return.

Let us now turn to the use of this model for investment and portfolio management. Portfolio managers can be divided into passive and active managers. Passive managers believe that mispriced securities cannot be identified and thus try to hold a portfolio that mimics some set of stocks. The most common way passive management is practiced is to hold a portfolio of stocks that closely tracks a selected index. Active management involves making bets about some securities or set of securities in the sense of designing a portfolio based on a belief that one or more securities is mispriced.

Passive Management

The multi-index model can play a major role in improving passive management. It can be used to do a better job of tracking an index or to design a passive portfolio that is appropriate for a particular client.

The simplest use of a multi-index model is to create a portfolio of stocks that closely tracks an index. An obvious way to construct an index fund is to hold stocks in the same proportion they represent of the index. However, many index funds do not simply hold each stock in an index in the proportion the stock represents of the index but rather attempt to replicate the index with a smaller number of stocks. The more issues in an index, the smaller the companies represented in an index, and the less liquid the stocks in an index, the more costly it is to match the index by purchasing stocks in the same proportion they represent in the index. Clearly, once one becomes concerned with tracking an index that represents a very large segment of a market, exact matching of proportions becomes less and less appropriate. An index fund can be created using the single-index model by finding the portfolio that has a beta of 1 with the desired index and that has minimum residual risks for a given portfolio size (minimum variance of the e_is in a single-index form of Equation [16.1]).

Employing a multi-index model rather than a one-index model allows the creation of an index fund that more closely matches the desired index.[18] The reason for this is clear. A properly constructed multi-index model ensures that the index has been matched in terms of all important sources of return movements (risk). On the other hand, just matching on market risk can leave the portfolio and the index with different sensitivities to the common factors affecting both, such as sensitivity to inflation. Let us consider a simple example of this. Reviewing the sensitivity coefficients associated with the market from Table 8.1, in Chapter 8, we see that both oil stocks and cyclical stocks have a sensitivity with the S&P index of 1.14. Thus, in a single-index model, except for residual risk, one would be indifferent to holding oil stocks or cyclical stocks in matching the S&P index. However, oil stocks and cyclical stocks have very different sensitivities (bs) to sales growth. Thus a portfolio that was matched to an index on sensitivity with the S&P but was not matched on the b value with sales growth might not track the index very well in periods when unexpected changes in sales growth were large.

In general, the fewer stocks in an index-matching portfolio, the less likely that the portfolio will be matched on the common factors affecting the portfolio and the index and the

[18]See Elton and Gruber (1988) for a demonstration of the improvement in index tracking that results from using a four-index as opposed to a one-index model.

greater the superiority of multi-index models over single-index models.[19] This is true because unexpected changes in the missing indexes will differentially impact the residual risk in future periods if sensitivity to these missing indexes is not held constant. Portfolios are often formed to serve as arbitrage portfolios in the trading of options or futures on an index. Firms typically attempt to form a small basket of stocks (25 or 50) that they can actively trade as they change their futures or options position. The number must be kept small, because the basket of stocks will be bought and sold frequently. The use of multi-index models becomes critical in these instances.

Another problem frequently encountered in passive management is the desire to match an index with a portfolio that excludes certain types of stocks. Social goals or management preferences frequently restrict the set of stocks that can be used to match an index. In the last 10 years, for example, it was not uncommon for a pension fund to declare that it would not own tobacco stocks or gambling stocks. It is likely that a sector of the market such as tobacco stocks has a sensitivity to inflation or interest rates that is different from that of the average stock. If an index fund is formed from a set of stocks that precludes tobacco stocks using the single-index model, then the sensitivity to the single index will be matched, but the sensitivity to other important influences will probably be different. Use of a multi-index model improves tracking an index.[20]

Multi-index models also help improve performance under a set of conditions that are directly opposite to those just described. An investor may decide to match an index with a portfolio that must contain certain stocks. This is very common in Japan, where stocks are often held for reasons that have their foundations in the business relationship between firms. In the United States, an investor may want to maintain (or add) certain holdings in a portfolio for business reasons or because the investor does not want to recognize certain accumulated but unrealized capital losses or gains for either tax purposes or reporting purposes.[21] The problem, then, is to find an overall portfolio matching as closely as possible an index but including a defined set of stocks. Because these stocks may have sensitivities to important influences that are different from the index being matched, it is important to explicitly match on each of the key risk factors.

One type of passive management that can be performed with a multi-index model is fundamentally different from what can be done with a single-index model. The multi-index model allows one to closely match an index while purposely taking positions with respect to certain types of risk different from the positions contained in the index. For example, consider a pension fund that has cash outflows affected by inflation (COLA or cost-of-living adjustments). The payments for such a pension fund increase with inflation. Thus the overseers want a portfolio that will perform especially well when the rate of inflation increases. This can be illustrated more fully by returning to the data presented for the S&P index earlier in this chapter. The b value (sensitivity) for the S&P index with inflation was -0.37, which implies (other things held constant) that an investment in the S&P index will tend to go down by 0.37% if the rate of inflation goes up by 1%. If a pension fund is particularly sensitive to inflation risk (because its liability payments go up with inflation), it might wish to hold a portfolio that has a zero sensitivity to inflation (or even a positive sensitivity). It could form a portfolio that had the same response to all factors affecting the S&P (except for the inflation

[19]See Elton and Gruber (1988) for empirical evidence on this issue.

[20]See Elton and Gruber (1988) for empirical evidence.

[21]An example of the latter occurs in insurance companies, where the realization of gains or losses impacts the surplus account and thus the ability of the firm to write new business.

factor) by solving a quadratic programming problem to form a portfolio that matched all S&P *b*s except for the *b* on inflation, had a zero or positive *b* with inflation, and had minimum residual risk.

The applications we have just discussed can be done using a multi-index model; however, assuming an APT adds additional insight into the process. It tells the investor the expected cost of changing the exposure to inflation. Observing the λ with inflation, we see that the market will accept a lower return of 4.32 for every one unit increase in sensitivity to inflation. This is because the aggregate of investors prefer stocks that offer higher return when inflation goes up. The investor who wanted zero sensitivity to inflation would expect to have a $(-4.32) \times 0.37 = -1.60$ change (decrease) in expected return to obtain the preferred position. Like most of economics, this is not a free lunch. Instead, it is a method of allowing the investor to make specified trade-offs between types of risk and expected returns.[22]

There is one variable in our model that allows the investor to take an action that is very close to a free lunch. Let us reexamine our model. One of the factors, oil price changes, had a zero λ (was given a zero price by the market). Although oil prices affect returns on some stocks, changes in oil prices are not a pervasive enough influence to be priced by the market. At first glance, one might think that the sensitivity on a portfolio to oil should be set to zero. After all, why take on a risk (increased variability in returns) with no commensurate increase in expected returns? For the average investor this is correct. However, think of an investor whose cash *outflow* increases with increases in oil prices. Such an investor would want to hold a portfolio of securities that has a positive sensitivity with oil prices. Furthermore, because oil sensitivity is not priced by the market, increasing the sensitivity to oil prices does not change expected return. Of course, if everybody wanted to hold portfolios that exhibited increased return with increases in oil prices, then the λ associated with oil prices would be positive. The fact that an investor desires, with respect to oil sensitivity, a position different from the aggregate allows the investor to improve his portfolio with no decrease in expected return, although there will be some increase in total risk.

Keep in mind that matching an index while making quantitative judgments on the amount of a particular type of risk to take can be done only if indexes representing these risks are contained in the multi-index model. Furthermore, the expected return (or expected cost) of these nonaverage risk positions can be determined from only an APT model.

Active Management

Most uses of multi-index models for active management parallel their use in passive management. It is easier to discuss them in reverse order to that presented previously. What a multi-index model does that cannot be done with a single-index model is allow the user to make factor bets. If you believe that unexpected inflation will accelerate at a rate above that anticipated by the market ($I_I > 0$), then you may want to place a bet by increasing your exposure (*b* value) with inflation. This can be done holding a portfolio with a sensitivity to inflation larger than the S&P index.[23] Obviously the more indexes included in the model, the more active bets you can make. For example, in the Salomon model described earlier in this chapter, you can take active bets on economic growth, the stage of the business

[22]Deviating from market *b*s to better match liabilities is different from deviating to take active bets on the change in one or more underlying influences. This active use of factor bets will be discussed shortly.

[23]We assume the S&P index is the relevant benchmark in this section. Actually the analysis holds with the sensitivities of any benchmark (growth stocks or the New York Stock Exchange index) substituted for sensitivities to the S&P index.

cycle, long-term interest rates, short-term interest rates, inflation rates, the value of the U.S. dollar, or the state of the stock market.

Return to the simple model we have been discussing and assume that the S&P index is the appropriate benchmark and that an analyst believed that sales were going to increase by 1% more than the market expected. The analyst might increase the *b* value with respect to sales on the portfolio from the 1.71 value found for the S&P index to 2.21. Under the APT model and recognizing that the λ for sales is 1.49, the increase in sales sensitivity of 0.5% would lead to a 0.5(1.49) = 0.745% increase in *expected return*, which is just sufficient to reward the investor for the additional risk. However, the additional 1% increase in sales would lead to an additional 2.21% increase in the return on the portfolio should it materialize. Of this 2.21% increase, 0.5% arose from increasing the sensitivity to sales, while 1.71 would have arisen had the *b* been left at the level of the S&P index. The 0.5% increase is often called the excess risk-adjusted return, which arises from an ability to forecast factors better than the market.

Multi-index models and APT models can be used just as the single-index model and CAPM models are used to form optimal portfolios building upon estimates of the performance of individual securities. The simplest approach is that discussed in Chapter 8, where a multi-index model is used to generate the covariance between securities while expected returns and variances are supplied by some combination of analysts' forecasts and historical data.

Another application of APT is to use APT to determine stocks that are under- or overvalued. In this procedure an analyst produces a forecast of the return on a stock. The APT is then used together with estimates of the sensitivity of the stock to the factors to calculate a required return for the stock (using an equation such as 16.10). If the estimated return is above what is required given the stock's sensitivity and the λs, the stock is purchased.

This is a generalization of the analysis that is used when the CAPM rather than the APT is used as an equilibrium model. Recall, as shown in Chapter 14, that the CAPM is a straight line in expected return beta space (see Figure 14.2). If a firm's expected return and beta are such that it plots above the CAPM line, it offers a higher return (given its beta) than is required in equilibrium and is a buy. Similarly, if it plots below the line, its expected return is less than required in equilibrium and it should be sold. The analysis with APT has the same logic. Consider a two-factor APT model. In this case, the APT plots as a plane in a three-dimensional space where the axes are sensitivities to the two factors and expected return. Firms that plot above the plane offer a higher expected return than is required given the sensitivities and λs and should be purchased.[24]

Why the APT rather than the CAPM? If the APT is the appropriate equilibrium model and the CAPM is used, then stocks with different sensitivities to the factors but the same market beta will be incorrectly classified as equally risky. The CAPM model incorrectly implies that they have the same expected return.

To better understand this, let us return to the example we have been discussing in this chapter. Note that the lambda on growth is positive. This implies that investors require a higher expected return for stocks that have higher sensitivity to unexpected changes in growth. A stock with a high sensitivity to growth will tend (because growth has a positive price or lambda) to have a higher expected return than a stock with a lower sensitivity to growth. But this is ignored (except for the part captured in the market beta) by the standard CAPM models. Thus the extra return investors require (as reflected in the market

[24]The market prices of risk (λs) for the CAPM or APT models can be specified by theory or estimated using analysts' forecasts of expected return and sensitivities for a set of stocks. See Chapter 19 for a discussion of how firm forecasts are used.

price of risk or lambda associated with high sensitivity to growth) will be interpreted as underpricing by the standard CAPM model. Stocks that are very sensitive to unexpected changes in growth will tend to lie above the security market line. Stocks that are sensitive to other priced influences not included in the single-index model are likely to show up as systematically underpriced or overpriced by the CAPM and to lie above or below the security market line.

One of the most common uses of the APT model is to form a portfolio of stocks that while closely tracking a target will also produce a return in excess of that index. One way to implement this type of procedure is simply to employ the index-matching procedure described earlier in this chapter but only allow selection from among a set of stocks that analysts have earmarked as superior performers. Other techniques use either numeric discrete ranking of stocks or expected return on stocks in an attempt to produce an excess return above an index, while using the multi-index model to track an index as closely as possible.[25] Portfolios designed this way have become known as *research-tilted index funds*. Although some additional risk is involved (the index cannot be matched as closely when selecting from a restricted set of stocks), investors who use this technique feel that an excess return can be earned with only a slight loss in the ability to track the index. The advantage of the multi-index model over the simple-index model is that the target index can be tracked more closely because the different sources of risk are explicitly taken into consideration.

The more the target being tracked differs from a diversified market portfolio, the more important it is to use a multi-index model. The extreme case and one that has received a lot of attention is the long-short investment strategy or risk-neutral strategy. If one has superior ability to identify stocks that will perform above average on an APT risk-adjusted basis and stocks that will perform below average on an APT risk-adjusted basis, then using the APT index, one can form portfolios that offer an excess return and have no risk (zero *b* risk) with respect to any factor (e.g., no risk due to change in the market level, inflation, or interest rate movements). Obviously there is also no expected return due to any factor because the beta on each factor is set to zero. What one gets is a pure payoff from security selection with all factors including the market neutralized. We can examine this by returning to Equation (16.11). If we believe that an analyst can predict the extra return from any security over a period of time (return from security selection), Equation (16.11) can be written as

$$R_i = R_F + \alpha_i + \lambda_1 b_{i1} + \lambda_2 b_{i2} + \lambda_3 b_{i3} + \cdots + b_{i1} I_1 + b_{i2} I_2 + b_{i3} I_3 + \cdots + e_i$$

where α_i is the extra return the security analyst predicts on security *i*.

Think of this equation for each of two portfolios: portfolio *L* is a portfolio of long positions and portfolio *S* is a portfolio of short positions. Furthermore, assume that the portfolios are formed so that $b_{Lj} + b_{Sj} = 0$ for all *j*s. Then, combining the preceding equation for each of the two portfolios, we get a risk-neutral (or, more specifically, a systematic-risk-neutral) portfolio denoted by *N* with a return given by

$$R_N = R_F + \alpha_L + \alpha_S$$

and with a risk given by

$$e_N = e_L + e_S$$

[25]Many firms have their analysts place stocks into groups (often five), with group 1 being the best purchases and group 5 the stocks that should be sold.

Burmeister, Roll, and Ross (1994) examined the payoff from such a model from the period April 1991 to March 1992, assuming αs could be correctly identified. They found that over this period the S&P index had a return of 11.57% per year and a standard deviation of 18.08%. Their factor-neutral portfolio had a return of 30.04% per year and a standard deviation of 6.26% per year. While these are obviously optimistic figures, for they assume perfect foresight, they do indicate the ability of factor-neutral portfolios to lower risk and, if forecasting ability exists, increase return.

Although one can perform the same type of analysis with a single-index model rather than a multi-index model, the overall risks of the portfolio will be greater and the user likely to find she is undertaking factor bets (inflation, interest rate, etc.) rather than pure security selection bets.

Factor Investing: An Active-Passive Approach

In this section we discuss a relatively new approach to investment management that focuses on capturing the premiums that result from exposure to systematic risk factors. We have seen that, in equilibrium, there is a positive expected rate of return in excess of the riskless rate associated with such things as pervasive market risk or exposure to other factors such as inflation risk. An investor who is less sensitive to these risks may choose to have a higher exposure to them in return for a higher expected premium. Such entities might be endowments with longer horizons or funds that are naturally "hedged" against certain factors such as a sovereign fund associated with an oil-producing country.

The inspiration for factor investing comes from the APT equation 16.7, which expresses the expected return of a security as

$$\overline{R_i = R_f + \sum_{j=1}^{j} b_{ij}\lambda_j}$$

where λ_j compensates the investor for bearing the risk of asset i's exposure to systematic risk factor j. The same model for a single security also explains the expected return of an entire portfolio. Investor i seeking returns above the riskless R_f can scale up expected excess returns by choosing a set of b_{ij} for the portfolio, such that the b_{ij} on high λ factors are high. The resulting portfolio is then simply an allocation across a set of factors with positive risk premia. In short, the APT can be a framework for taking risk as well as hedging away risk. Factor investing in this context does not seek to predict or "time" the variations in the factors but rather represents a strategic allocation across a set of factors depending on the investor's risk appetite for exposure to factor risks.[26]

In Chapter 8 we introduced fundamental multi-index models based on such things as industries, macroeconomic variables, security characteristics, and statistical factor extraction. Cochrane (1999) observes that a multifactor model of risk is entirely natural because the financial markets do not operate independently from the real economy. Investors have jobs that depend upon macroeconomic cycles, they confront the risks of inflation due to government policy, and they have occasional liquidity needs. These exposures logically impact their investment decisions as well and ultimately determine asset prices. The APT is extremely flexible in allowing a range of different factor models that capture these major sources of risks. Indeed, because so many factors in the economy interact with each other,

[26]Note that sometimes the term *factor investing* is applied to tactical asset allocation models that seek to "time" factor realizations.

it is unlikely that any factor model is unique. The key requirements, however, are, first, that the factors have positive expected premia and, second, that they represent pervasive, recognizable risks for which investors in aggregate demand compensation in the form of positive returns in excess of the riskless rate.

Different factor models for investment are used in industry, but the most academic models for the equity universe are those studied by Chan, Hamao, and Lakonishok (1991), Fama and French (1992), and Carhart (1997). These authors documented positive historical premia for long-short portfolios formed on characteristics of stocks, including market beta, firm capitalization, the book to market ratio, and momentum (that is, the relative rank of the stock return over the prior year). For example, a factor formed by sorting stocks each period by the book to market ratio, taking a long position in the 30% with a high book to market ratio and a short position in the bottom 30% with a low book to market, is referred to as the value factor, or HML (i.e., high minus low) for short. This portfolio is "active" in the sense that it requires regular rebalancing but passive in the sense that it follows a prespecified rule. Figure 16.1 plots the cumulative returns to the four Fama, French, and Carhart factors, ignoring transaction costs. Rm-Rf represents the equity risk premium, SMB is a small cap minus large cap factor, HML is the value factor, and MOM is a momentum factor. Notice that all four factor premia are significantly positive over the period from 1927 to 2012. What is also interesting is that they reacted differently to the financial crisis of 2008. The momentum factor experienced a much larger decline than did the size and value factors, although over the long term it had higher returns. This suggests that a combination of factors may provide some diversification in periods of distress. The graph also indicates, however, that the size and value factors performed much better over the earlier period. Whether this is due to changes in the risk premium demanded by

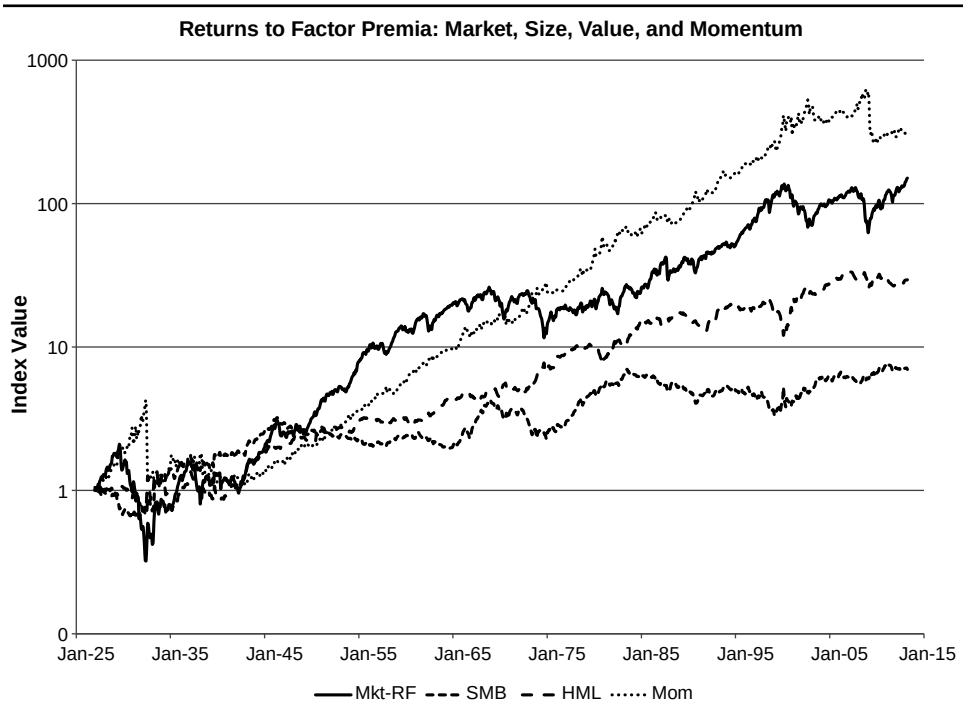


Figure 16.1 Fama, French, and Carhart Factor Performance, 1927–2012. *Source*: Data courtesy of Kenneth French.

investors to hold these portfolios or simply evidence that premiums shift through time is an open question.

Annualized Summary Statistics for Fama, French, and Carhart Factors, 1927–2011

	Geometric Mean (%)	Arithmetic Mean (%)	Standard Deviation (%)
Factor			
MOM	6.91	8.60	18.09
Rm-Rf	5.70	7.58	20.31
HML	3.30	4.13	13.42
SMB	2.54	3.16	11.57

The summary statistics for the factors show that, whereas the mean of the value and size factors is lower, their volatility is lower as well. This is because MOM, HML, and SMB are constructed from a long position in one portfolio of equities and a short position in another. We would thus expect them to be less volatile than a long-stock short Treasury bill portfolio like the equity premium Rm-Rf.

Ang and Kjaer (2011) and Blitz, Huij, and Steenkamp (2012), among others, argue that factor investing is a good long-horizon strategy, particularly for institutions that are not sensitive to occasional periods of poor returns. In a world where factor premia reflect aggregate aversion for exposure to the factor, an investor willing to accept factor risk may, over the long term, "harvest" the premia. Ang, Goetzmann, and Schaefer (2009) recommended a factor investing approach for the Norwegian sovereign wealth fund which invests for future generations of the Norwegian people. They propose a set of equity and fixed-income factors that have generated historical premia. The factor descriptions that follow are adapted from their 2009 report.

TERM STRUCTURE FACTOR

Long-term government bonds have historically provided higher yields than short-term bonds, and this difference is regarded as a compensation for the exposure to the risk in variation in the future short-term rate, although several theories of the yield curve propose additional reasons for this yield gap, including variation in demand for money at different maturities (cf. Vayanos and Vila, 2009). Embedded in the long-term rates are also expectations about inflation and inflation risk premiums, because long-term bonds are nominal securities (cf. Ang and Piazzesi, 2003). The term structure factor is formed by taking a long position in long-term government bonds and a short position in short-term Treasury securities. Chen, Roll, and Ross (1986) used this variable in their empirical test of the APT and found it to be a determinant of the cross section of stock returns, despite being a bond factor.

CREDIT RISK FACTOR

This captures the compensation for the risk of default on debt instruments. For risky corporate securities this is likely to be correlated to the equity premium, because defaulted debt becomes equity, and it also has a macroeconomic component, because defaults tend to be clustered in time and occur in periods of financiai distress. This factor is constructed from long positions in long-term corporate debt and short positions in corresponding long-term

government bonds of the same maturity or duration. Credit risk was also included in the original Chen, Roll, and Ross study.

FOREIGN EXCHANGE [FX] CARRY

This factor captures the return to lending in high-interest currencies and borrowing in low-interest currencies. This strategy has an implicit premium due to the risk of interest rate convergence, but its use has been documented only over the modern era for which currencies have traded in the capital markets, which is the post-1970s periods, after the breakdown of Bretton Woods. Jurek (2007) and many other authors document large gains over multiple year horizons for carry trade strategies but also point out that they are significantly negatively skewed, indicative of an insurance-like payoff. This factor goes long currencies with high yields and short currencies with low yields.

VALUE FACTOR

This is typically constructed from a long position in stocks with a high book to market ratio and a short position in stocks with an unusually low book to market ratio. In practice, many other indications of value may be used, including prices relative to other accounting variables (such as earnings, sales, forecasted and realized earnings). An economic interpretation of this factor is that it represents compensation for firm distress, because high book to market value is low market price. Other theories for about the value factor include the hypothesis that it compensates for low-growth options by inflexible firms with assets in place during periods of distress (cf. Zhang, 2005), or time-varying sensitivities of value stocks that manifest themselves as changing betas in macroeconomic states (cf. Ang and Chen, 2007). The behavioral explanation for the value premium is over extrapolation of past growth rates into the future (cf. Lakonishok, Shleifer, and Vishny, 1994).

SIZE FACTOR

Constructed from a long position in small cap stocks and a short position in large cap stocks, the small firm effect is well documented in financial economics literature, and yet the economics underlying the long-term outperformance of lower-capitalization stock is still not clearly understood. Researchers since Banz (1981) have explored a number of different risk-based, institutional, and behavioral explanations. Some of these theories are addressed in more detail in Chapter 17. Berk (1995) argues that the small firm effect is the result of a mispecified asset pricing model. The implication of his critique is that it may not be necessary to pin down the precise risks captured by small stocks because their small capitalization simply reflects exposure to one or more unidentified but nevertheless priced factors. This perspective on the small firm effect justifies the inclusion of small cap stocks in a factor portfolio.

MOMENTUM FACTOR

As with the value premium, this factor has a very strong historical premium but no clearly articulated risk. It is somewhat related to a strategy long used in practice called buying on relative strength. In the academic literature, Jegadeesh and Titman (1993) documented positive returns to buying past winners and selling past losers over the post-1926 period. Rouwenhorst (1998) observed profitable momentum returns in

international equity portfolios as well. Research by Chabot, Ghysels, and Jagannathan (2008) demonstrates that momentum existed in the Victorian era, indicating that it is not limited to a recent window in capital market history.

The most compelling explanations for momentum are behavioral and are based on investors underreacting to news (cf. Barberis, Shleifer, and Vishny, 1998). Cooper, Gutierrez, and Hameed (2004) and Chabot, Ghysels, and Jagannathan (2008) note that momentum profits depend on whether the stock market itself is in a bull or bear market. Momentum profits turn negative during an extended bear market—the implication being that bull markets attract naïve investors whose slow price equilibration may be exploited by simple investment rules. Figure 16.1 shows that the momentum factor crashed during the financial crisis, indicating that the steady profits it gained over long stretches of time may compensate for episodic but extreme risks. Momentum appears to be pervasive in nearly every asset class, within each asset class, and even across asset classes (cf. Asness, Moskowitz, and Pedersen, 2008). Most momentum factors have relatively high turnover.

VOLATILITY FACTOR

A volatility premium arises, among other reasons, because agents are averse to periods of increased volatility and are willing to pay high prices to hedge against significant increases in market volatility—which typically also coincide with downward market moves (cf. Bakshi and Kapadia, 2003). A volatility factor manifests itself in the cross section of options: out-of-the money options are expensive compared to at-the-money options (cf. Coval and Shumway, 2001), for example. In another formulation, differences in prices between options on indexes and individual options on index components—called correlation trades (cf. Driessen, Maenhout and Vilkov, 2007)—exploit this relation in several asset classes (fixed income, currencies, commodities, etc.). This factor is not restricted to just derivatives as any relation between volatility and returns should be captured by a volatility risk factor. For example, Ang, Goetzmann, and Schaefer (2009) show that stocks with low volatility have high returns in the global cross section of stock returns. Bollerslev, Tauchen, and Zhou (2009) show that high volatility premia formed from implied volatility measures forecast high future returns. Blitz and van Vleit (2007) show that options are not necessary to construct a volatility factor. By sorting global equities into portfolios based on historical volatility, they find a high premium to a factor that is long low-volatility stocks and short high-volatility stocks. Although the economic logic behind higher expected returns to low-volatility portfolios seems contrary, Asness, Frazzini, and Pederson (2011) argue that it is consistent with an aversion to leverage.

LIQUIDITY FACTOR

Liquidity, or the ability to trade a security quickly and with little impact on the market price, is a well-known risk in capital markets. The financial crisis in 2008 was characterized by a massive reduction in liquidity for certain instruments such as mortgage-backed securities. Even prior to the crisis, strong empirical evidence suggested that illiquidity arose in periods of financial distress.[27] Researchers and practitioners have hypothesized that illiquidity is therefore a priced factor that can deliver a premium to investors able to withstand illiquid episodes in the capital markets. David Swensen, chief investment officer of the Yale University Endowment, formally introduced the idea of a liquidity premium to investment practice in his

[27]Chordia, Sarkar, and Subrahmanyam (2005).

book *Pioneering Portfolio Management*. Over a period of two decades, the Yale Endowment allocated a large fraction to illiquid asset classes such as private equity partnerships.

Because Yale is a perpetually lived institution with a long investment horizon and less urgent liquidity needs, he reasoned that Yale could take a premium from counterparties in the capital market who relied on the need to liquidate their holdings during periods of distress. The long-term success of the Yale Endowment drew considerable attention to the liquidity premium strategy and many imitators.

Although the Yale Endowment successfully accessed a liquidity factor by holding illiquid, privately traded partnerships, considerable research using public capital market data suggests that an illiquidity premium factor can be formed by using public markets. Franzoni, Nowack, and Phalippou (2011) note that the factor exposure of private equity funds is correlated to liquidity factors measured in bond and stock markets. Early interest rate theories, for example, attributed the term *structure spread* to the value of cash immediacy. Ang, Goetzmann, and Shaefer (2009) construct a liquidity factor by forming a long-short position in on-the-run versus off-the-run Treasury bonds. They find that it explained a significant fraction of the movement in the fixed-income returns of the Norwegian sovereign wealth fund. Pastor and Stambaugh (2003) show that differences in liquidity explain differences in U.S. stocks and create a priced liquidity factor from equities. Chen, Ibbotson, and Hu (2010) construct investable portfolios of stocks to capture the liquidity premium and document a substantial premium. It is quite likely that liquidity is priced in most public capital markets and thus represents an opportunity for a patient investor, well insulated from urgent liquidity needs, to realize a substantial premium.

INFLATION FACTOR

Expectations of future inflation are embedded in the yield curve because most bonds—except TIPS—are contracts in nominal terms. Chen, Roll, and Ross (1986) found that two inflation-related factors were priced in the cross section of stock returns: inflation surprises and change in expected inflation. A long literature on inflation hedging documents that it is a key macroeconomic risk that investors seek to avoid and thus it must command a premium in expected security returns. Efforts to construct portfolios to track inflation risk include Lamont (2001) and Downing, Longstaff, and Rierson (2012). Because of the pervasive aversion to inflation, however, investing to "harvest" an inflation premium has not yet been proposed in practice. Chen, Roll, and Ross (1986) did not find a premium on exposure to oil price shocks—a proxy for inflation.

GDP FACTOR

Cochrane (1999) observes that the economic effects of booms and recessions in the economy are so pervasive that exposure to fluctuations in the GDP should command a risk premium. Chen, Roll, and Ross (1986) found a premium for a factor constructed from shocks to industrial production, despite the econometric difficulties in measuring macroeconomic variables. More recent tests have found some evidence that a procyclicality factor may explain differences in stock returns. Vassalou (2003), for example, finds that news about future realized changes in GDP is priced in the cross section of stock returns. Goetzmann, Watanabe, and Watanabe (2010) construct a variable from biannual economists' forecasts of GDP growth and find that a long-short portfolio based on procyclicality betas delivers a substantial risk premium. Campbell and Diebold (2009) show that economic forecasts of GDP growth predict changes in aggregate expected returns to the stock market, consistent with the existence of a pervasive GDP factor.

EQUITY RISK PREMIUM

Although discussed extensively elsewhere in this volume, it is important to include the equity risk premium as a factor like any other, if for no other reason than the fact that the spread of stock returns over bond returns has been empirically documented over centuries of stock market returns.[28]

LIMITATIONS OF FACTOR INVESTING

One important critique of factor investing is that the factors that deliver premia do not always have clear economic interpretations. Although some factors, such as credit and term structure risk, are easily linked to risks which investors naturally seek to avoid or insure themselves against, others, such as momentum, are puzzling. Unfortunately, the factors that appear to explain the most cross-sectional variation in historical stock returns are also those with the least economic intuition. Size, value, and momentum are very effective at explaining stock returns, and they have very significant historical premia. However, it has been difficult to attribute these premia to fundamental risks faced by investors. Although somewhat unsatisfying, the current state of knowledge about factors and their premia is likely to change with future financial research.

For investment managers, however, the uncertainty about the economic logic of these factors represents an important challenge, because if the source of the historical premiums is not well understood, then it may be difficult to reliably forecast the continuation of future premiums to such factors as momentum and value.

A second critique particularly relevant to practice is that the sum of the premia, while reliably linked to economic risks, is small in magnitude. For example, the yield curve premium is extremely reliable over long horizons, however, it rarely returns more than 3% per year. This is insufficient for a portfolio with a goal of generating substantial real returns.

To exploit smaller premiums as a source of excess returns, it may be necessary to use leverage to scale up the factor. Consider, for example, an equal-weighted portfolio of two factors: the yield-curve premium with a 2.5% annual return and the equity risk premium with an 8% annual return—values calculated from returns over the period 1926–2012. Their respective standard deviations are 20.66% and 8.48%. We measure the relative risk exposure of the two factors by their variances: 0.427% and 0.72%, respectively. Note that most of the variance of the portfolio is due to the equity premium, not the yield curve premium.

Risk parity, an approach advocated by some investment managers, such as Bridgewater, AQR, and Rebeco, argues that equalizing the risks of the two factors improves diversification and potentially increases expected returns. To see how this would work, choose a position for the yield curve factor such that its variance is equal to that of the equity risk premium: $W_{\text{yield curve}} = 0.427/0.72$. Assuming that the factor can be levered at the riskless rate by a factor of 5.93, this gives an expected return to the levered yield curve factor of 14.06%, while taking a risk (measured in variance or standard deviation) equal to the risk of the equity premium factor. Now a 50/50 portfolio of the levered yield curve factor and the unlevered equity premium factor is equal to 11.31%, with a volatility less than 20.66% because the two factors are not perfectly correlated.

The risk-parity approach increases expected return but also risk. Also, by scaling up one factor through leverage, it exposes the portfolio to additional risks such as liquidity.

[28]For overviews of long-term historical evidence on the equity premium, see Goetzmann and Ibbotson (2006) Dimson, Marsh, and Staunton (2008).

Leverage requires borrowing money, and borrowing requires cash resources to pay interest on the loan. In periods such as 2008, when borrowing became difficult, levered positions became difficult to maintain. Another critique of the application of risk-parity to factor investing is that mean–variance optimization will select the best allocation across the two factors. Risk parity is an arbitrary choice to maintain equal risk exposure—not the best choice. The mathematics of the Markowitz model insures that simply putting the factors into the program will dominate risk parity.

FACTOR INVESTING SUMMARY

In sum, factor investing is an important application of the Arbitrage Pricing Theory. It relies on a solid economic foundation for the source of profits based on compensation for taking systematic risks. The challenge of factor investing is identification of the factors and understanding the economics underlying the historical premia they have generated. Investment management in the factor investing framework allows measurement and monitoring of exposure to risk factors and the flexibility to "dial up" certain factors to suit the risk budget of the portfolio. Risk parity is one such proposed approach to guide the allocation across factors.

Performance Measurement and Attribution

The last use of multi-index and APT models we should examine is in the area of portfolio performance evaluation. It is difficult to discuss the use of APT in performance measurement and evaluation without reviewing the whole literature in this area. Because of this, we will leave a detailed discussion and the continuation of the example we started in this chapter until Chapter 25. However, consideration of the model we have discussed shows that the expected performance of any portfolio is not just a function of the portfolio's sensitivity to the market but also a function of the portfolio's sensitivity to sales growth and inflation. If influences that enter the return-generating process and APT are ignored in doing performance evaluation, not only cannot the analyst's performance be attributed to the type of management decisions he or she is making, but perhaps more important, incorrect conclusions may be reached about how well managers are performing.

CONCLUSION

In this chapter, we have reviewed

1. modern concepts of arbitrage pricing
2. alternative approaches to estimating arbitrage pricing models
3. some uses of arbitrage pricing models

Considerable evidence continues to be produced on the usefulness of arbitrage pricing models.

APPENDIX A

A SIMPLE EXAMPLE OF FACTOR ANALYSIS

To provide the reader who has never used any form of factor analysis with a demonstration of how it works, we include a simple example in this appendix. We choose to use

Table 16.2 Correlation Coefficient between Returns in Four Countries

	Belgium	France	Canada	U.S.
Belgium	1.0			
France	0.65	1.0		
Canada	0.38	0.41	1.0	
United States	0.41	0.43	0.72	1.0

principal component analysis for the example, because this leads to a solution that is easiest to interpret.[29]

We choose 10 years of monthly data on the Morgan Stanley Capital International stock indexes for each of four countries: the United States, Canada, France, and Belgium. Remember that principal components analysis extracts from these data the index that explains as much as possible of the correlation in returns between the four countries and then finds a second index that explains as much as possible of the correlation in returns not explained by the first index.[30] The indexes produced by principal components are formed by combining (weighting) the time series of return for each country with the mean return for each country extracted.

Before we perform principal component analysis, let us think about what we would expect the results to look like. We might hypothesize that the first index would be some sort of measure of how stocks in general did, that is, some general aggregation of the returns under study. In thinking about the problem, one would expect Canada and the United States to act somewhat alike and France and Belgium to act somewhat alike, whereas we would expect the differences between these paired countries to be greater. In fact, the correlations between the four countries as shown in Table 16.2 bear out this speculation.

The indexes that are the first two principal components estimated from these data are presented as follows. Remember that in performing principal components analysis, we do not specify the indexes we expect to find; we simply let the data determine the indexes.

The indexes are

$$I_{1t} = 0.67\left(R_{Bt} - \overline{R}_B\right) + 0.76\left(R_{Ft} - \overline{R}_F\right) + 0.76\left(R_{ct} - \overline{R}_c\right) + 0.77\left(R_{ut} - \overline{R}_u\right)$$

$$I_{2t} = -0.40\left(R_{Bt} - \overline{R}_B\right) - 0.37\left(R_{Ft} - \overline{R}_F\right) + 0.73\left(R_{ct} - \overline{R}_c\right) + 0.41\left(R_{ut} - \overline{R}_u\right)$$

where

I_{1t} and I_{2t} are the two indexes extracted from the data.

The *R*s are monthly returns, and the subscripts *B*, *C*, *F*, *U*, and *t* represent Belgium, Canada, France, the United States, and time.

Note that the first index is very close to an equally weighted index of all four markets and thus meets our expectation that the index that would explain as much as possible of returns is the general return index. The second index is long in North America and short in Europe. It meets our expectation that the second index should capture the fact that North American and European markets are less associated with each other than with markets within their own region.

To see how well these two indexes work, we can regress the returns from each country against the two indexes. When we did so, the R^2 were 0.81 for Belgium, 0.95 for Canada, 0.84 for France, and 0.74 for the United States.

[29]See Elton and Gruber (1994) for a detailed discussion on the use of factor analysis in multi-index models.

[30]Principal components then extract a third and fourth index. In this case, we report only the first two, because the third and fourth are not statistically significant.

APPENDIX B

SPECIFICATION OF THE APT WITH AN UNOBSERVED MARKET FACTOR

This appendix is a brief recapping of the procedures put forth in a series of articles by Burmeister, McElroy, and others. For further details see articles by Burmeister and others (1987, 1988, 1986, 1988).

We can represent a return-generating process (multi-index model) with observable indexes plus an unobservable index designated by index *k* as

$$R_{it} = \overline{R}_{it} + \sum_{j=1}^{J} b_{ij} F_{jt} + b_{ik} F_{kt} + \epsilon_{it} \tag{B.1}$$

Making the no-arbitrage assumption of APT, expected return is approximately given by

$$\overline{R}_{it} = \lambda_{Ot} + \sum_{j=1}^{J} b_{ij} \lambda_{jt} + b_{ik} \lambda_{kt} \tag{B.2}$$

We will make the assumption of McElroy and Butmeister (1988) that all λ_{ot} equal the risk-free rate and all other λs are constant over time, substituting (B.2) into (B.1):

$$R_{it} = R_{Ft} + \sum_{j=1}^{J} b_{ij} \lambda_j + \sum_{j=1}^{J} b_{ij} F_{jt} + b_{ik} \lambda_k + b_{ik} F_{kt} + \epsilon_{it} \tag{B.3}$$

Now assume a very well-diversified portfolio called *m*. For this portfolio, residual risk approaches zero and[31]

$$R_{mt} = \lambda_m + R_{Ft} + \sum_{j=1}^{J} b_{mj} F_{jt} + F_{kt} \tag{B.4}$$

$$\text{where } \lambda_m = \sum_{j=1}^{J} b_{mj} \lambda_j + \lambda_k$$

Burmeister and McElroy assume the market portfolio has no residual risk $\epsilon_{mt} = 0$; this F_{kt} is the unobserved error term.

However, F_{kt} can be estimated by the residual of an ordinary least squares (OLS) time-series regression of R_{mt} on the observed variables as in Equation (B.4), or rearranging, (B.4) yields

$$\hat{F}_{kt} = \left(R_{mt} - R_{Ft}\right) - \left[\lambda_m + \sum_{j=1}^{J} b_{mj} F_{jt}\right] \tag{B.5}$$

McElroy and Burmeister (1988) show that $\hat{F}_{kt}$ is an unbiased estimate of the common stocks F_{kt}. Thus substituting $\hat{F}_{kt}$ for F_{kt} in (B.3) and adjusting the residual yields

$$R_{it} = R_{Ft} + \sum_{j=1}^{J} b_{ij} \lambda_j + b_{ik} \lambda_k + \sum_{j=1}^{J} b_{ij} F_{jt} + b_{ik} \hat{F}_{kt} + e_{it} \tag{B.6}$$

[31]The assumption is made that the unobserved variable F_{kt} is recalled to have a beta of 1 with the portfolio *m*.

To estimate this equation, McElroy and Burmeiser (1988) first use time series analysis to estimate Equations (B.4 and B.5) and then nonlinear, seemingly unrelated regressions to estimate (B.6).[32]

In doing so, they make one more interesting change in this model. They allow for the possibility that although APT correctly prices every security in their sample, it may not correctly price every security in the highly diversified portfolio.

QUESTIONS AND PROBLEMS

1. Assume that the following two-index model describes returns:

$$R_i = a_i + b_{i1}I_1 + b_{i2}I_2 + e_i$$

Assume that the following three portfolios are observed.

Portfolio	Expected Return	b_{i1}	b_{i2}
A	12.0	1	0.5
B	13.4	3	0.2
C	12.0	3	−0.5

Find the equation of the plane that must describe equilibrium returns.

2. Referring to the results of Problem 1, illustrate the arbitrage opportunities that would exist if a portfolio called *D* with the following properties were observed:

$$\overline{R}_D = 10 \quad b_{D1} = 2 \quad b_{D2} = 0$$

3. Repeat Problem 1 if the three portfolios observed have the following characteristics.

Portfolio	Expected Return	b_{i1}	b_{i2}
A	12	1.0	1
B	13	1.5	2
C	17	0.5	−3

4. Referring to the results of Problem 3, illustrate the arbitrage opportunities that would exist if a portfolio called *D* with the following characteristics were observed:

$$\overline{R}_D = 15 \quad b_{D1} = 1 \quad b_{D2} = 0$$

5. If we accept the Sharpe model as a description of expected returns, using the data in Table 16.1, find the expected return on a stock in the construction industry with the following characteristics. Assume a riskless rate of 8%:

Beta = 1.2
Yield = 6
Size = 0.4
Bond beta = 0.2
Alpha = 1

[32]Conditions on the relationship between e_{it} and ϵ_{it} are delineated in McElroy and Burmeister (1988).

6. Return to Problem 1. If $(\bar{R}_m - R_F) = 4$, find the values for the following variables that would make the expected returns from Problem 1 consistent with equilibrium determined by the simple (Sharpe–Lintner–Mossin) CAPM:

A. $\beta_{\lambda 1}$ and $\beta_{\lambda 2}$

B. β_p for each of the three portfolios

C. R_F

BIBLIOGRAPHY

1. Admati, Anat, and Pfleiderer, Paul. "Interpreting the Factor Risk Premia in Arbitrage Pricing Theory," *Journal of Economic Theory*, **35** (Feb. 1985), pp. 191–195.
2. Ang, Andrew, Chen, Joseph, and Xing, Yuhang. "Downside Risk and the Momentum Effect," NBER working paper 8643 (2001), http://www.nber.org/papers/w8643.
3. Ang, A., and Piazzesi, M. "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables," *Journal of Monetary Economics* **50** (2003), pp. 745–787.
4. Ang, A., Chen, J., and Xing, Y. "Downside Risk," *Review of Financial Studies* **19** (2006), pp. 1191–1239.
5. Ang, Andrew, Goetzmann, William, and Schaefer, Stephen. "Evaluation of Active Management of the Norwegian Government Pension Fund—Global—," Norwegian Ministry of Finance (2009), http://www.regjeringen.no/.
6. Ang, A., Hodrick, R. J., Xing, Y., and Zhang, X. "High Idiosyncratic Volatility and Low Returns: International and Further U.S. Evidence," *Journal of Financial Economics* **91** (2009), pp. 1–23.
7. Ang, Andrew, and Kjaer, Knut N. "Investing for the Long Run" (2011). Available at SSRN.
8. Blitz, David, Huij, Joop, and Steenkamp, Tom. *Factor Investing* (Robeco Institutional Management, 2012).
9. Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. "Value and Momentum Everywhere," working paper, NYU (2008).
10. Asness, Cliff, Frazzini, Andrea, and Pedersen, Lasse H. "Leverage Aversion and Risk Parity," *Financial Analysts Journal* **68,** No. 1 (2012), pp. 47–59.
11. Bakshi, G., and Kapadia, N. "Delta-Hedged Gains and the Negative Market Volatility Risk Premium," *Review of Financial Studies* **16** (2003), pp. 527–566.
12. Barberis, N., Shleifer, A., and Vishny, R. W. "A Model of Investor Sentiment," *Journal of Financial Economics* **49** (1998), pp. 307–343.
13. Berk, Jonathan B. "A Critique of Size-Related Anomalies," *Review of Financial Studies* 8, No. 2 (1995): pp. 275–286.
14. Berry, Michael, Burmeister, Edwin, and McElroy, Marjorie. "Sorting Out Risks Using Known APT Factors," *Financial Analysts Journal* (March 1988), pp. 29–42.
15. Black, Fisher, Jensen, Nick, and Scholes, Myron. "The Capital Asset Pricing Model: Some Empirical Tests," in M. Jensen (ed.), *Studies in the Theory of Capital Markets* (New York: Praeger, 1972).
16. Blake, Christopher, Elton, Edwin J., and Gruber, Martin J. "The Performance of Bond Mutual Funds," *Journal of Business*, **66,** No. 3 (July 1993), pp. 371–403.
17. Blitz, David, and Vliet, Pim Van. "The Volatility Effect: Lower Risk without Lower Return," *Journal of Portfolio Management* (2007), pp. 102–113.
18. Bollerslev, Tim, Tauchen, George, and Zhou, Hao. "Expected Stock Returns and Variance Risk Premia," *Review of Financial Studies*, **22,** No. 11 (2009), pp. 4463–4492.
19. Bower, Dorothy H., Bower, Richard S., and Logue, Dennis E. "Arbitrage Pricing Theory and Utility Stock Returns," *Journal of Finance*, **39,** No. 4 (Sept. 1984), pp. 1041–1054.

20. Brennan, M. "Capital Asset Pricing and the Structure of Security Returns," working paper, University of British Columbia (1971).
21. ——. "Discussion," *Journal of Finance*, **36** (May 1981), pp. 352–357.
22. Brown, S. J., and Weinstein, M. I. "A New Approach to Testing Asset Pricing Models: The Bilinear Paradigm," *Journal of Finance*, **38,** No. 3 (June 1983), pp. 711–743.
23. Burmeister, Edwin, and McElroy, Marjorie. "APT and Multifactor Asset Pricing Models with Measured and Unobserved Factors: Theoretical and Econometric Issues," discussion paper, Department of Economics, University of Virginia and Duke University (1987).
24. Burmeister, Edwin, and McElroy, Marjorie. "Joint Estimation of Factor Sensitivities and Risk Premia for the Arbitrage Pricing Theory," *Journal of Finance*, **43,** No. 3 (July 1988), pp. 721–733.
25. Burmeister, Edwin, and Wall, Kent. "The Arbitrage Pricing Theory and Macroeconomic Factor Measures," *The Financial Review*, **21** (Feb. 1986), pp. 1–20 .
26. Burmeister, Edwin, Roll, Richard, and Ross, Stephen, "A Practitioner's Guide to Arbitrage Pricing Theory," in *A Practitioner's Guide to Factor Models* (Charlottesville, VA: The Research Foundation of the Institute of Chartered Financial Analysts, 1994).
27. Burmeister, Edwin, Wall, Kent, and Hamilton, James. "Estimation of Unobserved Expected Monthly Inflation Using Kalman Filtering," *Journal of Business and Economic Statistics*, **4** (April 1986), pp. 147–160.
28. Campben, Sean D., and Diebold, Francis X. "Stock Returns and Expected Business Conditions: Half a Century of Direct Evidence," J*ournal of Business and Economic Statistics*, **27,** No. 2 (2009), pp. 266–278.
29. Carhart, Mark M. "On Persistence in Mutual Fund Performance," J*ournal of Finance*, **52,** No. 1 (1997), pp. 57–82.
30. Chabot, B., Ghysels, E., and Jagannathan, R. "Price Momentum in Stocks; Insights from Victorian Age Data," NBER working paper 14500 (2008).
31. Chan, K. C., Chen, Nai-fu, and Hsiech, David. "An Explanatory Investigation of the Firm Size Effect," *Journal of Financial Economics*, **14** (Sept. 1985), pp. 451–471.
32. Chan, Louis K. C., Hamao, Yasushi, and Lakonishok, Josef. "Fundamentals and Stock Returns in Japan." *Journal of Finance*, **46,** No. 5 (1991), pp. 1739–1764.
33. Chen, N. "The Arbitrage Pricing Theory: Estimation and Applications," working paper, Graduate School of Management, UCLA (1981).
34. Chen, Nai-fu. "Some Empirical Tests of the Theory of Arbitrage Pricing," *Journal of Finance*, **38,** No. 5 (Dec. 1983), pp. 1393–1414.
35. Chen, Nai-fu, and Ingersoll, Jonathan E., Jr. "Exact Pricing in Linear Factor Models with Finitely Many Assets: A Note," *Journal of Finance*, **38,** No. 3 (June 1983), pp. 985–988.
36. Chen, Nai-fu, Roll, Richard, and Ross, Stephen. "Economic Forces and the Stock Market," *Journal of Business*, **59** (July 1986), pp. 386–403.
37. Chen, Zhiwu, Ibbotson, Roger, and Hu, Wendy. "Liquidity as an Investment Style," (2010).
38. Cho, D. Chinhyung. "On Testing the Arbitrage Pricing Theory: Inter-Battery Factor Analysis," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1485–1502.
39. ——. "Some Fundamental Factors Effecting Asset Prices," working paper, University of Wisconsin, (1984).
40. Cho, D. Chinhyung, and Taylor, William. "The Seasonal Stability of the Factor Structure of Stock Returns," *Journal of Finance*, **42** (Dec. 1987), pp. 1195–1211.
41. Cho, D. Chinhyung, Elton, Edwin J., and Gruber, Martin J. "On the Robustness of the Roll and Ross Arbitrage Pricing Theory," *Journal of Financial and Quantitative Analysis*, **XIX,** No. 1 (March 1984), pp. 1–10.
42. Cho, D. Chinhyung, Eun, Cheol S., and Senbet, Lemma W. "International Arbitrage Pricing Theory: An Empirical Investigation," *Journal of Finance*, **41,** No. 2 (June 1986), pp. 313–329.
43. Chordia, Tarun, Sarkar, Asani, and Subrahmanyam, Avanidhar. "An Empirical Analysis of Stock and Bond Market Liquidity," *Review of Financial Studies*, **18,** No. 1 (2005), pp. 85–129.
44. Cochrane, John H. "Production-Based Asset Pricing and the Link between Stock Returns and Economic Fluctuations," *Journal of Finance*, **46,** No. 1 (March 1991), pp. 209–237.

45. Cochrane, John H. "New Facts in Finance," working paper w7169, National Bureau of Economic Research (1999).
46. Connor, G. "A Factor Pricing Theory for Capital Assets," working paper, Kellog Graduate School of Management, Northwestern University (1981).
47. Connor, Gregory. "A Unified Beta Pricing Theory," *Journal of Economic Theory*, **34,** No. 3 (Oct. 1984), pp. 13–31.
48. Connor, G., and Korajczyk, R. "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, **15,** No. 3 (1986), pp. 373–394.
49. Conway, Delores, and Reinganum, Marc. "Capital Market Factor Structure: Identification through Cross Validation," *Journal of Business and Financial Statistics*, **6,** No. 1 (Jan. 1988), pp. 1–15.
50. Cooper, M. J., Gutierrez, R. C., and Hameed, A. "Market States and Momentum," *Journal of Finance*, **59** (2004), pp. 1345–1365.
51. Coval, J. D., and Shumway, T. "Expected Option Returns," *Journal of Finance*, **56** (2001), pp. 983–1009.
52. Daniel, Kent. "Momentum Crashes," Columbia Business School Research Paper 11–03 (2011).
53. Davis, J. L., Fama, E. F., and French, K. R. "Characteristics, Covariances, and Average Returns: 1929 to 1997," *Journal of Finance*, **55** (2000), pp. 389–406.
54. Dhrymes, Pheobus J., Friend, Irwin, and Gultekin, N. Bulent. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory," *Journal of Finance*, **39,** No. 2 (June 1984), pp. 323–346.
55. Dimson, E., Marsh, P., and Staunton, M. *Triumph of the Optimists: 101 Years of Global Investment Returns* (Princeton, NJ: Princeton University Press, 2002).
56. Dimson, E., Marsh, P., and Staunton, M. *Triumph of the Optimists: 101 Years of Global Investment Returns* (Princeton, NJ: Princeton University Press, 2008).
57. Downing, Christopher T., Longstaff, Francis A., and Rierson, Michael A. "Inflation Tracking Portfolios," working paper w18135, National Bureau of Economic Research (2012).
58. Driessen, J., Maenhout, P., and Vilkov, G. "The Price of Correlation Risk: Evidence from Equity Options," *Journal of Finance* (forthcoming).
59. Dybvig, Phillip H. "An Explicit Bound on Deviations from APT Pricing in a Finite Economy," *Journal of Financial Economics*, **12** (1983), pp. 483–496.
60. Elton, E., and Gruber, M. "Non-Standard CAPMs and the Market Portfolio," *Journal of Finance*, **39,** No. 3 (1984), pp. 911–924.
61. Elton, Edwin J., and Gruber, Martin J. "A Multi-index Risk Model of the Japanese Stock Market," *Japan and the World Economy*, **1,** No. 1 (1988), pp. 21–44.
62. Elton, Edwin J., and Gruber, Martin J. "Expectational Data and Japanese Stock Prices," *Japan and the World Economy*, **1** (1989), pp. 391–401.
63. Elton, Edwin J., and Gruber, Martin J. "Multi-Index Models Using Simultaneous Estimation of All Parameters," in *A Practitioner's Guide to Factor Models* (Charlottesville, VA: The Research Foundation of the Institute of Chartered Financial Analysts, 1994).
64. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher. "Fundamental Variables, APT, and Bond Fund Performance," *Journal of Finance*, **50,** No. 4 (1995), pp. 1229–1256.
65. Elton, Edwin J., Gruber, Martin J., and Mei, Jianping. "Cost of Capital Using Arbitrage Pricing Theory: A Case Study of Nine New York Utilities," *Journal of Financial Markets, Institutions & Instruments*, **3,** No. 3 (1994), pp. 1–37.
66. Elton, Edwin J., Gruber, Martin J., and Rentzler, Joel. "The Arbitrage Pricing Model and Returns on Assets under Uncertain Inflation," *Journal of Finance*, **38,** No. 2 (May 1983), pp. 525–538.
67. Fama, Eugene. "Stock Returns, Real Activity, Inflation and Money," *American Economic Review*, **71** (1981), pp. 545–565.
68. Fama, Eugene, and French, Kenneth. "The Cross Section of Expected Stock Returns," *Journal of Finance*, **47,** No. 2 (June 1992), pp. 427–466.
69. Fama, Eugene, and French, Kenneth. "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, **33,** No. 1 (2003), pp. 3–56.

70. Fama, Eugene, and French, Kenneth. "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, **51,** No. 1 (1996), pp. 55–84.
71. Fama, Eugene and French, Kenneth. "Profitability, Investment and Average Returns," *Journal of Financial Economics*, **82 (**2006), pp. 491–518.
72. Fama, Eugene, and Gibbons, Michael. "A Comparison of Inflation Forecasts," *Journal of Monetary Economics*, **13** (1984), pp. 327–348.
73. Fama, Eugene, and MacBeth, James. "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, **38** (1973), pp. 607–636.
74. Franzoni, Francesco, Nowak, Eric, and Phalippou, Ludovic. "Private Equity Performance and Liquidity Risk," Swiss Finance Institute Research Paper 09–43 (2011).
75. Frazzini, Andrea, and Pedersen, Lasse. "Betting against Beta," Swiss Finance Institute Research Paper 12–17 (2011), http:/ssrn.com/abstract=2049939.
76. Garman, Mark B., and Ohlson, James A. "A Dynamic Equilibrium for the Ross Arbitrage Model," *The Journal of Finance*, **35,** No. 3 (June 1980), pp. 675–684.
77. Gibbons, M. "Multivariate Tests of Financial Models: A New Approach," *Journal of Financial Economics*, **10,** No. 1 (March 1982), pp. 3–27.
78. Gibbons, M. R. "Empirical Examination of the Return Generating Process of the Arbitrage Pricing Theory," working paper, Stanford University (1981).
79. Gibbons, Michael, Ross, Stephen, and Shanken, Jay. "A Test of the Efficiency of a Given Portfolio," *Econometrica*, **57,** No. 5 (1989), pp. 1121–1152.
80. Goetzmann, William N., and Ibbotson, Roger G. *The Equity Risk Premium: Essays and Explorations*. (New York: Oxford University Press, 2006).
81. Goetzmann, W. N., Watanabe, A., and Watanabe, M. "Investor Expectations, Business Conditions, and the Pricing of Beta-Instability Risk," working paper, SSRN (2009).
82. Goetzmann, William N., Watanabe, Akiko, and Watanabe, Masahiro. "Procyclical Stocks Earn Higher Returns," working paper, Yale School of Management (2010).
83. Grinblatt, Mark, and Titman, Sheridan. "Factor Pricing in a Finite Economy," *Journal of Financial Economics*, **12** (1983), pp. 497–507.
84. ——. "Approximate Factor Structures: Interpretations and Implications for Empirical Tests," *Journal of Finance*, **40** (1985), pp. 1367–1373.
85. ——. "The Relation between Mean–Variance Efficiency and Arbitrage Pricing," *Journal of Business*, **60** (1987), pp. 97–113.
86. Grinold, Richard, and Kahn, Ronald. "Multi-Factor Models for Portfolio Risk," in *A Practitioner's Guide to Factor Models* (Charlottesville, VA: The Research Foundation of the Institute of Chartered Financial Analysts, 1994).
87. Hansen, Lars, and Singleton, Kenneth. "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Assets Returns," *Journal of Political Economy*, **91** (1983), pp. 249–265.
88. Harvey, C. R., and Siddique, A. "Conditional Skewness in Asset Pricing Tests," *Journal of Finance*, **55** (2000), pp.1263–1295.
89. Harman, H. *Modern Factor Analysis*, 3rd ed. (Chicago: University of Chicago Press, 1976).
90. Huberman, Gur. "A Simple Approach to Arbitrage Pricing Theory," *Journal of Economic Theory*, **78** (1982), pp. 183–191.
91. ——. "A Review of the Arbitrage Pricing Theory," in John, Eatwell, Murray, Milgate, and Peter, Newman, (eds.), *The New Palgrave: A Dictionary of Economic Theory and Doctrine* (New York: Stockton Press, 1987).
92. Huberman, Gur, and Kandel, Shmuel. "Mean–Variance Spanning," *Journal of Finance*, **42,** No. 4 (Sept. 1987), pp. 873–888.
93. ——. "Mean–Variance Spanning," *Journal of Finance*, **42** (Sept. 1987), pp. 873–888.
94. Huberman, Gur, and Stambaugh, Robert. "Mimicking Portfolios and Exact Arbitrage Pricing," *Journal of Finance*, **42** (March 1987), pp. 1–9.
95. Huberman, Gur, Kandel, Shmuel, and Stambaugh, Robert F. "Mimicking Portfolios and Exact Arbitrage Pricing," *Journal of Finance*, **42,** No. 1 (March 1987), pp. 1–9.
96. Hughes, P. "A Test of the Arbitrage Pricing Theory," working paper, University of British Columbia (1981).

97. Ibbotson, Roger, and Sinquefield, Rex. *Stocks, Bonds, Bills and Inflation: The Past and the Future* (Charlottesville, VA: Financial Analysts Research Foundation, 1982).
98. Ikeda, Shinsuke. "Arbitrage Asset Pricing under Exchange Risk," *Journal of Finance*, **46,** No. 1 (March 1991), pp. 447–455.
99. Ingersoll, Jonathan E., Jr. "Some Results in the Theory of Arbitrage Pricing," *Journal of Finance*, **39** (1984), pp. 1021–1039.
100. ——. *Theory of Financial Decision Making* (Totowa, NJ: Rowman and Littlefield, 1987).
101. Jegadeesh, N., and Titman, S. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, **48** (1993), pp. 65–91.
102. Jobson, J.D. "A Multivariate Linear Regression Test for the Arbitrage Pricing Theory," *Journal of Finance*, **37,** No. 4 (Sept. 1982), pp. 1037–1042.
103. Joreskog, K.G. *Statistical Estimation in Factor Analysis* (Stockholm: Almqvist & Wiksell, 1963).
104. Joreskog, K.G. "Some Contributions to Maximum Likelihood Factor Analysis," *Psychometrika*, **32,** No. 4 (Dec. 1967), pp. 443–482.
105. Joreskog, K.G. "Factor Analysis by Least Squares and Maximum Likelihood Methods," in K. Enslein, A. Ralston, and H. S. Wilf (eds.), *Statistical Methods of Digital Computers* (New York: John Wiley & Sons, 1977).
106. Jurek, J. W., "Crash-Neutral Currency Carry Trades," working paper, Princeton University (2007).
107. King, B. "Market and Industry Factors in Stock Price Behavior," *Journal of Business*, **39** (Jan. 1966), pp. 139–190.
108. Kogan, Leonid, and Tian, Mary. "Firm Characteristics and Empirical Factor Models: A Data-Mining Experiment," discussion paper, FRB International Finance (Nov. 28, 2012), http://ssrn.com/abstract=218213.
109. Kryzanowski, L., and To, M.C. "General Factor Models and the Structure of Security Returns," *Journal of Financial and Quantitative Analysis*, **18,** No. 1 (March 1983), pp. 31–37.
110. Lakonishok, J., Shleifer, A., and Vishny, R. W. "Contrarian Investment, Extrapolation, and Risk," *Journal of Finance* **49** (1994), pp. 1541–1578.
111. Lamont, Owen A. "Economic Tracking Portfolios," *Journal of Econometrics*, **105** (2001), pp. 161–184.
112. Lawley, D.N. "The Estimation of Factor Loadings by the Method of Maximum Likelihood," *Proceedings of the Royal Society of Edinburgh, Section A*, **60** (1940), pp. 64–82.
113. Lawley, D.N., and Maxwell, M.A. *Factor Analysis as a Statistical Method* (London, UK: Butterworths, 1963).
114. Lehmann, Bruce, and Modest, David. "The Empirical Foundations of the Arbitrage Pricing Theory I: The Empirical Tests," *Journal of Financial Economics*, **21** (1988), pp. 213–254.
115. Lettau, Martin, and Ludvigson, Sydney. "Resurrecting the (C)CAPM: A Cross-Sectional Test When Risk Premia Are Time-Varying," *Journal of Political Economy*, **109,** No. 6 (2001), pp. 1238–1287.
116. Levine, M.S. *Canonical Analysis and Factor Comparison* (Beverly Hills, CA: Sage Publications, 1977).
117. Lewellen, Jonathan, Nagel, Stefan, and Shanken, Jay. "A Sceptical Appraisal of Asset Pricing Tests," *Journal of Financial Economics*, **96,** No. 2 (2010), pp. 175–194.
118. Li, Qing, Vassalou, Maria, and Xing, Yuhang. "Sector Investment Growth Rates and the Cross Section of Equity Returns," *Journal of Business*, **79,** No. 3 (2006), pp. 1637–1665.
119. Lintner, J. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Review of Economics and Statistics*, **47** (Feb. 1965), pp. 13–37.
120. Litzenberger, R. H., and Ramaswamy, K. "The Effect of Personal Taxes and Dividends on Capital Asset Prices: Theory and Empirical Evidence," *Journal of Financial Economics*, **7** (1979), pp. 163–196.
121. Lucas, Robert E., Jr. "Asset Prices in an Exchange Economy," *Econometrica*, **46** (1978), pp. 1429–1445.

122. Lustig, Hanno, and van Nieuwerburgh, Stijn. "Housing Collateral, Consumption Insurance, and Risk Premia: An Empirical Perspective," *Journal of Finance*, **60,** No. 3 (2005), pp. 1167–1219.
123. McElroy, Marjorie, and Burmeister, Edwin. "Arbitrage Pricing Theory as a Restricted Nonlinear Multivariate Regression Model: ITNLSUR Estimates," *Journal of Business and Economic Statistics*, **VI,** No. 1 (Jan. 1988), pp. 29–42.
124. McElroy, Marjorie, and Wall, Kent. "Two Estimators for the APT Model When Factors Are Measured," *Economics Letters*, **19** (1985), pp. 271–275.
125. Merton, Robert C. "An Intertemporal Capital Asset Pricing Model," *Econometrica*, **41** (1973), pp. 867–887.
126. Mossin, J. "Equilibrium in a Capital Asset Market," *Econometrica*, **34** (Oct. 1966), pp. 768–783.
127. Nagel, Stefan. "Empirical Cross-Sectional Asset Pricing," working paper, SSRN (Nov. 13, 2012), http://ssrn.com/abstract=217530.
128. Pástor, L'uboš, and Stambaugh, Robert F. "Liquidity Risk and Expected Stock Returns," *Journal of Political Economy*, **111,** No. 3 (2003), pp. 642–685.
129. Pástor, L'uboš. "Comparing Asset Pricing Models: An Investment Perspective," *Journal of Financial Economics*, **56,** No. 3 (June 2000), p. 335.
130. Reinganum, M. "The Arbitrage Pricing Theory: Some Empirical Results," *Journal of Finance*, **36** (May 1981), pp. 313–321.
131. Roll, R. "A Critique of the Asset Pricing Theory's Tests," *Journal of Financial Economics*, **4** (May 1977), pp. 129–176.
132. Roll, R. "Ambiguity When Performance Is Measured by the Securities Market Line," *Journal of Finance*, **33** (Sept. 1978), pp. 1051–1069.
133. Roll, R., and Ross, S. A. "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance*, **35,** No. 5 (Dec. 1980), pp. 1073–1103.
134. Roll, Richard, and Ross, Stephen A. "A Critical Reexamination of the Empirical Evidence on the Arbitrage Pricing Theory: A Reply," *Journal of Finance*, **39,** No. 2 (June 1984), pp. 347–350.
135. Rosenberg, Barr, Reid, Kenneth, and Lanstein, Ronald. "Persuasive Evidence of Market Inefficiency," *Journal of Portfolio Management* **11** (1985), pp. 9–17.
136. Ross, S.A. "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, **13** (Dec. 1976), pp. 341–360.
137. ——. "Return Risk, and Arbitrage," in Irwin Friend and James L. Bicksler (eds.), *Risk and Return in Finance*, Vol. 1 (Cambridge, MA: Ballinger, 1977).
138. Santos, Tano, and Veronesi, Pietro. "Labor Income and Predictable Stock Returns," *Review of Financial Studies*, **19,** No. 1 (2006), pp. 1–44.
139. Shanken, J. "The Arbitrage Pricing Theory: Is It Testable?" *Journal of Finance*, **37,** No. 5 (Dec. 1982), pp. 1129–1140.
140. ——. "Multi-Beta CAPM or Equilibrium-APT? A Reply," *Journal of Finance*, **40** (1985), pp. 1186–1189.
141. Shanken, Jay. "On the Estimation of Beta-Pricing Models," *Review of Financial Studies*, **5,** No. 1 (1992), p. 1SS.
142. ——. "Multivariate Tests of the Zero-Beta CAPM," *Journal of Financial Economics*, **14** (Sept. 1985), pp. 327–348.
143. Sharpe, W. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, **19** (Sept. 1964), pp. 425–442.
144. ——. "Factors in NYSE Security Returns, 1931–1979," *Journal of Portfolio Management*, **8,** No. 2 (Summer 1982), pp. 5–19.
145. Sinclair, N.A. "Security Return Data and 'Blind' Factor Analysis," working paper, Australian Graduate School of Management (1981).
146. Solnik, Bruno. "International Arbitrage Pricing Theory," *Journal of Finance*, **38,** No. 2 (May 1983), pp. 449–458.
147. Sorensen, Eric, Mezrich, Joseph, and Thum, Chee. "The Salomon Brothers U.S. Stock Risk Attribute Model," (Salomon Brothers, Oct. 1989).

148. Sorensen, Eric, Salomon, R. S., Davenport, Caroline, and Fiore, Maria. "Risk Analysis: The Effect of Key Macroeconomic and Market Factors on Portfolio Returns," (Salomon Brothers, Nov. 1989).
149. Stambaugh, Robert. "On the Exclusion of Assets from Tests of the Two-Parameter Model," *Journal of Financial Economics*, **10** (Nov. 1982), pp. 237–268.
150. ——. "Testing the CAPM with Broader Market Indexes: A Problem of Mean Deficiency," *Journal of Banking and Finance*, **7** (March 1985), pp. 5–16.
151. Swensen, David F. *Pioneering Portfolio Management: An Unconventional Approach to Institutional Investment* (New York: Simon and Schuster, 2009).
152. Tiemann, Jonathan. "Exact Arbitrage Pricing and the Minimum-Variance Frontier," *Journal of Finance*, **43,** No. 2 (June 1988), pp. 327–338.
153. Trzcinka, Charles. "On the Number of Factors in the Arbitrage Pricing Model," *Journal of Finance*, **41,** No. 2 (June 1986), pp. 347–368.
154. Vayanos, D., and Vila, J. L. "A Preferred-Habitat Model of the Term Structure of Interest Rates," working paper, LSE (2009).
155. Vuolteenaho, T. "What Drives Firm-Level Stock Returns?" *Journal of Finance*, **57,** No. 1 (2002), pp. 233–264.
156. Yogo, Motohiro. "A Consumption-Based Explanation of Expected Stock Returns," *Journal of Finance*, **61,** No. 2 (2006), pp. 539–580.
157. Zhang, L. "The Value Premium," *Journal of Finance*, **60** (2005), pp. 67–103.

Part 4

SECURITY ANALYSIS AND PORTFOLIO THEORY

17

Efficient Markets

The Efficient Market Hypothesis (EMH) is a theory that the price of a security reflects all currently available information about its economic value. A market in which prices fully reflect all available information is said to be *efficient*. The concept is important for investment management because it serves as a guide to expectations about the potential for profitable trading, the likelihood of finding an investment manager who can beat the market, and the limits of predictability in the capital markets. If the theory is precisely true, it is impossible for a speculator, an investment manager, or the clients of the manager to consistently beat the market.

The intuition underlying the EMH is the invisible hand of the marketplace. In a quest for profits, competition among speculators to buy undervalued assets or sell overvalued assets will quickly drive expected gains to trade to zero. The statement that prices "reflect all available information" implies that no trader has any kind of informational advantage in the security markets. If this is so, then the price today reflects the common or "market" expectation of what the security would be worth tomorrow.

The formal theoretical expressions of the EMH actually do not imply that prices are set in some kind of competitive market equilibrium. Nor do they specify the mechanism by which prices "reflect all available information." More significantly, the theory does not imply that market prices are "right," or that market expectations are formed in some rational way. However, the theory is often popularly interpreted to mean that markets are entirely rational. In fact, although a competitive equilibrium in security prices does imply an efficient market, interestingly enough, the reverse is not necessarily true.

In a purely speculative market the only reason to trade is to make a profit on special information. The resulting change in price would contradict the existence of a competitive equilibrium. It follows then that in an equilibrium where no one has any reason to trade, the market price of each security reflects the common or market information shared by all investors. In this chapter we review the theory and tests of the EMH with a focus on implications for portfolio management. The theory has undergone considerable change as researchers have sought to model real-world market mechanisms, costly information, and the role of agency.

EARLY DEVELOPMENT

Because of the importance of the hypothesis to investment practice, it has generated a vast literature extending over 150 years. One of the earliest formal expressions of market efficiency is a book written in 1863 by Jules Regnault, a broker on the Paris Bourse and an amateur mathematician.[1] He argued by force of logic and personal observation that the market price of a security at any given time reflected the wisdom of the crowd. In rather strong terms he claimed that speculators trading on market imperfections were delusional—that the only way to profit was to trade on private information that no one else had. He used probability theory to estimate the "gamblers ruin" problem: the number of trades until an uninformed speculator would lose all of his money.

Regnault was also the first person to argue that market efficiency implied that asset prices should follow a random walk. He tested this theory with historical French and British bond data and was thus the first empirical researcher to document a "random walk" in security prices. In 1900, the French mathematician Louis Bachelier (1900), in his doctoral thesis on option prices trading on the Paris Bourse, based his model of security price movements on a more rigorous expression of Regnault's random walk. Independently, and prior to Albert Einstein, Bachelier developed the equations of Brownian motion—a continuous-time expression of a random process—which relied on the implicit assumption of unpredictable price movements. His basic insight ultimately led to option pricing models in current use today.

After Bachelier, tests of the random walk hypothesis and theories about random security selection and random price behavior dominated thinking about the EMH for much of the twentieth century. For example, the investor and philanthropist Alfred Cowles (1933) studied whether professional market forecasters could beat random stock picks. His result: they did no better than chance. In a follow-up paper, Cowles and Jones (1937) tested the random walk model on U.S. stock prices.

In the postwar period, renewed interest in efficient markets stimulated asset pricing theory. The mathematician (and father of fractal geometry) Benoit Mandelbrot (1963) derived the random walk hypothesis in a general framework allowing for discontinuities and extreme events. Two years later, Nobel laureate Paul Samuelson published a famous paper, "A Proof That Properly Anticipated Prices Fluctuate Randomly." In it, he noted that the EMH implies only that "the market quotation . . . already contains in itself all that can be known about the future and in that sense has discounted future contingencies as much as humanly possible." In short, futures prices should be unbiased based on information available at the time prices are established, and speculation should be a "fair game" with an expected reward of zero or, more generally, an amount that reflects a normal risk premium. That same year, at the University of Chicago, Eugene Fama (1965) formalized the argument using the law of iterated expectations. Over time, Fama has been the main academic "guru" of the EMH. Much of the research on efficient markets since that time has relied on Eugene Fama's articles and insights.

However, the Random Walk Hypothesis developed by Bachelier and examined empirically by Kendall (1953) and other early statisticians is quite a bit more restrictive than the EMH articulated by Fama and his students. The Random Walk Hypothesis states that increments in (the logarithm) of prices should be independently distributed with fixed and finite variance, while the EMH merely states that the current stock prices

[1]Regnault (1863). See Jovanovic and Le Gall (2001).

reflect all available information. Indeed, events of the financial crisis of 2007 show that the idea of fixed and finite variance of stock market price movements is somewhat unrealistic. The confusion between the Random Walk Hypothesis (which, among other things, implies that the serial covariance of stock market price movements should be zero) and Fama's EMH permeates much of the empirical literature on this topic.

THE NEXT STAGES OF THEORY

One of the dominant themes in the academic literature since the 1960s has been the concept of an efficient capital market.[2] Although the reader may well be able to visualize several meanings of the term *efficient market*, and although it has, in fact, been used to denote different phenomena at different times, it has come to have a very specific meaning in finance. When someone refers to efficient capital markets, she means that *security prices fully reflect all available information*.

This is a very strong hypothesis. A necessary condition for investors to have an incentive to trade until the prices fully reflect all the information is that the cost of information acquisition and trading be zero. Because these costs are clearly positive, a more realistic definition is that prices reflect information until the marginal costs of obtaining information and trading no longer exceed the marginal benefit. Throughout this chapter when we are reviewing the evidence on market efficiency, some deviations from efficient markets will be observed. We will often comment on the likely size of transaction costs. However, the ultimate judgment in deciding if these deviations exceed reasonable transaction costs will be left to the reader.

Some authors require that prices accurately reflect fundamental information for a market to be efficient. However, most tests of the EMH simply deal with how fast information is incorporated but do not deal with whether it is correctly incorporated in prices. We will refer to the hypothesis that prices reflect fundamental values as market rationality and discuss these tests at the end of the chapter.

The EMH has historically been subdivided into three categories, each dealing with a different type of information. *Weak-form* tests are tests of whether all information contained in historical prices is fully reflected in current prices. *Semistrong-form* tests of the EMH are tests of whether publicly available information is fully reflected in current stock prices. Finally, *strong-form* tests of the EMH are tests of whether all information, public or private, is fully reflected in security prices and whether any type of investor can make an excess profit.[3]

These classifications were originally suggested by Fama (1988). In a recent review article Fama expanded the definition of the first type of efficiency. He changed the classification *weak-form tests* to the more general category *tests of return predictability*. We will adopt this generalization. Under this classification, we will examine patterns in security returns such as high returns in January and on Mondays as well as whether returns can be predicted from past data. Consistent with this new classification, Fama has changed *semistrong-form efficiencies* to *event studies* or *studies of announcements*, and we will also adopt this classification.

---

[2]This chapter benefited greatly from the review article by Fama (1991).

[3]Our definition of strong-form tests is different from that contained in the literature. Fama defines strong-form tests as tests of whether markets fully reflect nonpublic information. This is examined by analyzing whether any group of investors can earn excess returns. We believe that if excess returns were found, the tests could not differentiate as to whether the excess returns arose from monopoly access to information or superior use of publicly available information. Thus a more general definition of strong-form efficiency than Fama's is necessary.

Careful consideration will show that much of the efficient market literature is actually concerned with the speed with which information is impounded into security prices. For example, assume a firm announces that earnings will be three times larger than expected next year, with no additional investment on the part of the firm. Furthermore, suppose that there have been fundamental changes in the company that imply that this increase in the level of earnings is permanent. Finally, assume that investors believe this announcement. Clearly the company is worth considerably more than before. The share price should go up to reflect this increase in value. The EMH does not deny the usefulness of this information, nor does it deny that prices should increase. What the EMH is concerned with is under what conditions an investor can earn excess returns on this security. Consider several scenarios.

First, assume that after the announcement, the price gradually increases over the week in response to the announcement. Investors examining the price sequence would observe that the price was moving away from that level at which it had previously traded. If they purchased securities when the securities started to trade away from historical prices, they would purchase the security a day or two after the announcement (after they had observed this new price behavior). If it took a week for the price to fully reflect the announcement, however, investors purchasing securities on the basis of movements away from historical prices would benefit from part of the price increase and make excess returns. Tests of the predictability of returns (formerly tests of the weak form of the EMH) are in part tests of whether this type of trading behavior can lead to excess profits. If returns are not predictable from past returns, then new information is incorporated in the security price sufficiently fast that, by the time an investor could tell from the price movements themselves that there had been a fundamental change in company prospects, the fundamental change is already fully reflected in price.

Consider a second scenario. Assume the investor hears the announcement of the improved prospects and believes it. The investor immediately buys shares of the company in anticipation of a price rise. The semistrong-form tests of the EMH are tests of whether this strategy leads to excess profits. The semistrong form of the EMH assumes that investors who wish to sell the security, as well as those who wish to buy, hear the announcement and reassess the value of the security. This reassessment leads to an immediate increase in price. The new price need not be the new equilibrium price, but it is not systematically lower or higher than the equilibrium price.[4] Thus an investor who buys the security after the announcement may be paying too little or too much for the security. If the semistrong form of the EMH holds, then over a large number of similar situations the investor would be paying on average about what the securities are worth. The investor would be unable to earn an excess profit by purchasing securities on the basis of such announcements.

The strong form of the EMH is concerned with two different ideas. Both can be demonstrated in terms of our previous example. One idea involves whether anyone can earn money by acting on the basis of information such as the announcement discussed earlier. Tests of the semistrong form of the EMH would examine all announcements such as the one under discussion, assume an investor purchased immediately after the announcement, and see if this leads to excess returns. There is nothing in this type of test that considers the value of the information contained in the announcement. Assume the investor hears

[4]It may take several days or weeks before investors can fully assess the impact of the change in firm conditions. Thus the price may be very volatile for a number of days. In efficient markets, the price immediately after the announcement is an unbiased estimate of the equilibrium after investors have fully assessed the impact of the earnings increase.

the announcement and can fairly accurately reassess its effect on the value of the company. When the price after the announcement is below the reassessed value, the investor purchases; when it is above, the investor sells if the shares are owned or shorts the stock (or does nothing) if the shares are not owned. The strong form of the EMH states that there is no investor with this superior ability. Because it is impossible to determine exactly how investors might utilize the announcement to reassess the value of the firm, tests of the strong form of the EMH are examinations of whether an investor or groups of investors have earned excess returns. Because of the lack of data on most types of investors, the group most frequently tested is managers of mutual funds.

The strong form of the EMH has a second facet that can also be illustrated with this example. Suppose the managers of the firm knew about the improved prospects in advance of the announcement; they had access to the information before it was publicly available. Could they purchase the security on the basis of the private information and make money? The most extreme form of the strong form of the EMH says no. It should not surprise the reader that the evidence does not support this extreme form of the EMH. What might surprise the reader, initially, is the strength of the evidence in favor of the less extreme forms. Once the reader considers the ideas behind these hypotheses, however, it should not be as surprising. Information about securities is rapidly disseminated. There are thousands of people who follow securities professionally. Information should be rapidly incorporated in price.

RECENT THEORY[5]

While information is an essential part of the EMH, most early theoretical models assumed that information was costless to obtain. None of them specified how information is generated or how trading actually impounded that information into prices. This presents a paradox. Why should a speculator do any research if trading on it is unprofitable? Yet, without speculators, how can prices impound information? Grossman and Stiglitz (1976) address this paradox in a model where investors pay for information. In their model speculators invest in research and recoup the cost of their investment by marginally profitable trades which, in turn, push prices toward fair economic value. The Grossman–Stiglitz world is a market driven by informed, active research, and speculation.

In the same year that Grossman and Stiglitz published their paper, Stephen Ross introduced the Arbitrage Pricing Theory (APT), based on the existence, or the possibility, of arbitrageurs exploiting mispriced systematic factor exposures in securities. These two theories of asset prices introduced the realism of human agents, active trading, and information production into the EMH, however, they also relied upon some additional financial assumptions about the process of arbitrage. In particular, while the APT allows for uniformed, irrational, suboptimal traders in the economy, it relies on the existence of a rational marginal investor to ultimately set prices. The rational arbitrageurs need not even be large or wealthy. However, a crucial assumption is that at least one of the rational arbitrageurs needs unlimited ability to borrow cash or short stocks to drive away mispricing. This raises the question, what if such financing were difficult?

In 1997, Shleifer and Vishny explored the arbitrageur's problem in a paper titled "Limits of Arbitrage." In their model, arbitrageurs face financing constraints because creditors have a short horizon. If mispriced assets do not converge quickly enough to

[5]This section adapted from Ang et al. (2010).

economic value, the arbitrageur becomes insolvent. In the Shleifer–Vishny world, market inefficiencies can persist when financing risk is high. Ironically, the paper preceded by one year the collapse of Long-Term Capital Management (LTCM), a large, highly levered hedge fund. Among other things, the fund bet on the convergence of U.S., European, and Japanese bond yields following the Asian currency crisis. Yields eventually converged, but not before LTCM was forced to liquidate. The Shleifer–Vishny paper highlights a fundamental conflict between rational arbitrageurs and sentiment-driven traders in the economy whose actions push assets away from economic value. Their model shows how the capital structure and institutional framework for arbitrage matter. Their view of the markets is not necessarily at variance with the EMH. If market expectations are driven by sentiment, then security prices would reflect these sentiment-driven factors. Their work has led researchers to try to identify behavior as well as fundamental factors in security returns.

The EMH has strong implications for security analysis. If, for example, empirical tests find that future return cannot be predicted from past return, then trading rules based on an examination of the sequence of past prices are worthless. If the semistrong form of the hypothesis is supported by empirical evidence, then trading rules based on publicly available information are suspect. Finally, if the strong-form tests were to show efficiency, then the value of security analysis itself would be suspect. Thus an understanding of efficient market tests should provide guidance for the reader in determining what types of analysis are useful.

This chapter is divided into five sections. The first section provides some additional background on the EMH. The next three sections discuss efficient market tests, and the last section discusses market rationality.

SOME BACKGROUND

To test any of the three forms of the EMH, it is necessary to be a little more precise regarding terms such as *excess return*. The purpose of this section is to introduce some of the terminology of the efficient market literature.

The discussion in the previous section is consistent with the process determining prices being a "fair game." *Fair game* is a very descriptive term. It says that there is no way to use "information" available at a point of time (t) to earn a return above normal. To clarify this further, let ϕ_t represent a set of information that is available to investors at a time t. Now, based on this information, the investor can make an estimate of what a stock's return will be between time t and time $t + 1$. The investor can then compare the estimated return with the equilibrium return. Perhaps the estimate of equilibrium return comes out of one of the models discussed in Chapters 13 and 16. Deviations of the investor's estimated return from the equilibrium return should contain no information about future returns. Whether the investor's estimate of return is above or below equilibrium should be unrelated to whether actual return is above or below equilibrium. There is no way the investor can use the information in the set ϕ_t to make a profit beyond that which is consistent with the risk inherent in the security.

This discussion may seem either intuitively obvious or completely unappealing. To further clarify, let us specify some conditions under which it would not be correct. Let us assume the information set ϕ_t contains real information that is not incorporated in stock price at time t but that will be incorporated at time $t + 1$. For example, assume that a government employee in charge of military contracts is about to approve a large contract for a small and previously unused supplier of butter to the army. This contract will result in a huge increase in profit for the company, but the market has assessed the

probability of the company getting it as very small. Thus only a fraction of the potential profits is incorporated in price. The procurement officer could make a much larger return than the equilibrium return for this company by purchasing its stock. The fair game model would not hold with respect to him. Thus, if the information set available to an investor is not incorporated in price, the fair game model does not hold with respect to that information set.

For the fair game model to hold, there must be no way in which the information set ϕ_t can be used to earn above equilibrium returns. For tests of return predictability, ϕ_t is defined as the past history of stock prices, company characteristics, market characteristics, and the time of the year. For semistrong tests, it is defined as the announcement of one or more pieces of information. For strong-form tests, it is defined as all information, whether publicly available or not, that is at the disposal of some group of investors.

The reader should note that there is no implication in any of our discussion that the expected return on any security is zero. One would expect that, in general, it would not only be different from zero but, in fact, be positive. Furthermore, one would expect that the return is related to risk with the more risky securities offering the higher return.

The reader might well wonder why we bother mentioning such an obvious point. This point has been a source of great confusion to many writers. One frequently reads that if the EMH holds, then the best estimate of tomorrow's price is today's price, or an expected return of zero. This is not a correct implication of the efficient market model. Rather, the implication is that the past information contains nothing about the magnitude of the deviation of today's return from expected return.

Before leaving this section, one additional term should be introduced—the *random walk model*. The random walk model assumes that successive returns are independent and that the returns are identically distributed over time. To understand the random walk model, visualize a roulette wheel with various returns written on it. Each period, the wheel is spun, and the return for the next period is read from the wheel. The outcomes from spins of the wheel are unrelated through time so that past returns are unrelated to future returns. Furthermore, the same wheel is spun each period, which causes the returns to be identically distributed.

The random walk model is a restricted version of the fair game model discussed earlier. The fair game model does not require identical return distributions in the various periods. Furthermore, the fair game model does not imply that returns are independent through time. For example, a firm could be increasing its debt and risk over successive periods of time and show increasing expected and increasing actual returns. In this case we would observe a correlation in the sequence of returns and past returns that could be used to predict future returns. However, because risk is increasing, and therefore expected return is also increasing, this information could not be used to earn an excess return. If the random walk hypothesis holds, the EMH must hold with respect to past returns (though not vice versa). Thus evidence supporting the random walk model is evidence supporting efficiency with respect to past returns.

TESTING THE EMH[6]

Tests of the EMH fall into two broad categories: price studies and manager studies. Studies of prices have generally focused on a search for trading rules that generate positive risk-adjusted investment returns when back-tested on historical data. Because of the widespread acceptance of the EMH in the latter part of the last century, any such rule reported

[6]Adapted from Ang et al. (2010).

in the academic literature was labeled an "anomaly"—a puzzle that challenged theory and could not be explained by economic theory. We discuss some of these in this next section. The anomaly studies share certain drawbacks, however.

First, because they use past return data, the actual conditions under which past prices were generated cannot be replicated, nor can the quality of historical price information ever be perfectly validated. Second, testing for whether a trading rule generated past risk-adjusted returns requires a definition of risk adjustment. This means that all tests of market efficiency are actually joint tests of efficiency and the model used for risk adjustment.[7] Thus a test rejecting market efficiency might simply be due to the failure of the researcher's model to correctly specify a risk adjustment factor. Finally, most tests of market efficiency are essentially tests of the law of one price, which posits that, in an efficient market, two economically equivalent assets will have the same value. However, the definition of "economically equivalent" may range from having exactly the same cash flows, to having the same systematic risk exposures, to having the same expected value. The power of the test thus depends on the reliability of the definition of economic equivalency.

The second broad category of efficiency tests are studies of investment managers that test the ability of active managers to generate risk-adjusted returns. The discovery that money managers as a group outperform a random or passive strategy of investing would contradict the main implication of the EMH. It is entirely possible that researchers could fail to find predictability in security prices, but reject the EMH when using managed portfolios. This would simply suggest that professional money managers are better at finding rejections of the law of one price than academic researchers. The tests of manager skill are of course subject to the same drawbacks as price studies, but in addition they are also joint tests of economic theories of agency. That is, it may be possible that professional managers exploit market inefficiencies, but the profits are not passed along to their customers because of incentive or monitoring problems inherent in the agency relationship between managers and customers.

TESTS OF RETURN PREDICTABILITY

In this section we review the studies examining the predictability of return from past data. In the first section we examine seasonal patterns in returns. A number of studies find that returns are different depending on the day of the week or time of the year. In the second section we discuss the predictability of return using past return. We analyze both short-term predictability and long-term predictability. In the third section we examine return and firm characteristics. In particular, we discuss evidence that abnormal returns are associated with small firms, firms with low market-to-book ratios, and firms with low earnings-to-price ratios. Finally, we discuss research showing a relationship between average firm or market characteristics and long-run return.

TESTS ON PRICES AND RETURNS

One of the earliest empirical challenges to the EMH is the existence of seasonal patterns in returns. As one of the early researchers on seasonality put it in 1924,

> Seasonal variations of security prices are impossible. . . . If a seasonal variation in stock prices did exist, general knowledge of its existence would put an end to it.[8]

[7]Cf. Brown and Warner (1985).
[8]Owens and Hardy (1925).

Despite this prediction, in 1942, Sidney Wachtel found a robust pattern of higher returns for U.S. stocks around the turn of the year—a pattern he conjectured might be explained by tax-loss selling and repurchase or by behavioral factors such as optimism and holiday cheer.[9] This pattern came to be known as the "January Effect" and has persisted since its original discovery, despite being widely known.[10] Since Wachtel's discovery, a number of researchers have found other seasonal patterns in security returns. Lakonishok and Smidt (1988) provide an excellent review of these seasonal anomalies and a test of many of them out of sample. In general, returns to the stock market, or sectors of the stock market, display some variation depending on the time of the day, the day of the week, the month of the year, the season of the year, and the condition of the weather, and even lunar cycles.[11]

One possible explanation for these discoveries of seasonalities of all sorts is data-mining. With hundreds of researchers examining the same set of stock returns, they are bound to find patterns simply by chance. If this is true, then evidence from other markets and other time periods should not find similar patterns—however many of the basic seasonality patterns have been replicated out of sample.[12] A second possible explanation is that these patterns are induced by the market structure and order flow.[13] The third possible answer is that markets are inefficient because one would expect that the patterns would disappear as investors exploited them. The best advice we can give the reader is that in most cases, because of transaction costs, the return differences are not large enough to develop a trading strategy to take advantage of them; if one is trading anyway, however, one might time the trade to try to exploit the pattern.

Intraday and Day-of-the-Week Patterns One pattern that has been extensively examined is the difference in return for various days of the week. Returns on Mondays are much lower than on other days of the week on the New York Stock Exchange (NYSE). Gibbons and Hess (1981) examined the 17-year period 1962–1978. They found that Monday's return was a negative –33.5% on an annualized basis. Furthermore, when they split the data into two subperiods, 1962–1970 and 1970–1978, the same large negative Monday return occurred. Gibbons and Hess also report a large positive return on Wednesdays and Fridays. In a more recent study Harris (1986) examined intraday and day-of-the-week patterns for the 14-month period from December 1981 to January 1983. He confirmed the large negative Monday return but found returns on the other four days to be positive and of roughly the same order of magnitude. The larger negative Monday return was not evenly spread during the day. Rather, half of it occurred between Friday's market close and Monday's open: the weekend return. Of the remaining decline, most occurred within the first 45 minutes of trading on Monday. After the first 45 minutes, returns on Monday closely resembled returns on any other day. On all days he found prices rose in the last 30 minutes of the day.[14] To date, no one has demonstrated

[9]Wachtel (1942).

[10]Ciccone (2011).

[11]I.e., a time-of-the-day effect (Harris, 1986); a day-of-the-week effect (cf. Ball and Bowers, 1986; Cross, 1973; French, 1980; Gibbons and Hess, 1981; Jaffe and Westerfield, 1985; Keim and Stambaugh, 1984; Lakoniskok and Levi, 1982); a time-of-the-month effect (Ariel, 1987); a season-of-the-year effect (Kamstra, Kramer, and Levi, 2003); a weather effect (Hirshleifer and Shumway, 2003); and lunar cycles (Yuan, Zheng, and Zhu, 2006).

[12]An interesting contrary result on seasonality is obtained for the markets in the nineteenth century, however. Cf. Zhang and Jacobsen (2012).

[13]Cf. Goetzmann, and Zhu (2005), who find that the U.S. weather efffect is explained by seasonality in bid–ask spreads.

[14]Keim (1989) finds that there is some tendency for Friday's closing prices to be at the ask rather than the bid. This would make Monday's prices somewhat lower even if there was no change in the bid and ask. However, Keim and Stambaugh (1984) still find the weekend effect after accounting for this.

Table 17.1 January Effect: 1926–2012

Months	S&P	Small	Corp	Govt
1	1.475	7.495	0.829	0.090
2	0.320	1.480	0.007	0.288
3	0.151	−0.280	0.190	0.334
4	1.381	1.221	0.210	0.449
5	0.109	0.091	0.271	0.193
6	1.450	1.045	0.375	0.521
7	2.450	3.001	0.245	0.172
8	1.859	2.133	0.213	−0.035
9	−0.861	−0.624	0.255	−0.026
10	0.021	−0.922	0.357	0.439
11	1.281	1.157	0.246	0.441
12	1.551	0.803	0.571	0.272
Average	0.932	1.383	0.314	0.262
not January	0.883	0.828	0.267	0.277
stddev.s	4.939	9.165	1.707	1.617
2 × stderr	0.533	0.988	0.184	0.174
t-value	1.111	6.746	3.054	−1.076

profitable trading strategies based on these patterns. However, the result suggests an investor should sell late Friday and purchase on Monday after the first 45 minutes. As usual, the reader should be cautioned that the study covers a short period of time, and the market may have adjusted to these patterns.

MONTHLY PATTERNS

As discussed previously, extensive research finds that returns in January are substantially higher than returns in other months. This is especially true for small stocks. Table 17.1 is an updated summary of the January effect using data from 1926 through 2012. It reports the mean monthly return for return indexes of four asset classes: large-cap stocks represented by the S&P 500, small-cap stocks represented by the smallest quintile of the NYSE by size, corporate bonds, and long-term government bonds. The average with and without January included is reported, as well as the sample standard deviation, the standard error of the January mean return, and a *t*-test of equality of the January mean and the annual mean excluding January. Notice that the mean for large-cap stocks is not statistically significant, but the effect for small-cap stocks is quite strong. The table also suggests that there is a July–August effect as well.

The January effect has been studied abroad as well as in the United States. Gultekin and Gultekin (1983) studied January return patterns in 17 countries including the United States. They found much higher returns in January than in non-January months for all the countries they studied. In fact, for the period they studied, the effect was bigger in the 16 non-U.S. markets. Kato and Shallheim (1985) examined excess returns in January and the relationship between size and the January effect for the Tokyo stock exchange. They found no relationship between size and return in non-January months. However, they found excess returns in January and a strong relationship between return and size, with the smallest firms returning 8% and the largest less than 3%.

Interestingly, there is a January effect in corporate bonds as well.[15] Table 17.1 shows a significant return for corporate bonds and none for government bonds. Keim and Stambaugh (1986) studied returns in bond markets from 1926 to 1978. They found that, on average, only in January do lower-quality bonds give an extra return.

Keim (1989) offers a microstructure explanation for part of the January effect. The CRSP tape calculates returns by using the closing price each month or the average of the bid and ask if the stock did not trade. Keim found that the last trade in December was primarily at the bid, which causes the return to appear high in the first few days of January. For example, assume a stock was 20 bid $20\frac{1}{4}$ ask. The last trade in December was likely to be at 20, whereas the first trade in January was somewhere between 20 and $20\frac{1}{4}$, on average $20\frac{1}{8}$. Thus, even without a change in the bid and ask, computing return using trading prices would imply a return of $\frac{1}{8}/20$ per day, or a very large annual return.

Keim found that the tendency for stocks to be at the bid price for the last trade in December was much more pronounced for small stocks. In addition, small stocks have a higher bid–ask spread and a lower price. Therefore the effect would be bigger for small stocks and would partly explain the differences in the January effect between large and small stocks. Thus part of the January effect can be explained by the prices having a tendency to be at the bid in December.

A second explanation that has been offered for the high returns in January (especially in the first few days of January) is a tax-selling hypothesis. A popular suggestion of investment advisers, at year-end, is to sell securities for which an investor has incurred substantial losses before the end of the year and purchase an equivalent security. This creates a tax loss for the investor. If the tax loss is substantial, it should more than cover transaction costs. Since the selling is in late December and the purchasing in early January, the argument is that prices are depressed at the end of December and rebound in January, creating high returns in January.

Both Reinganum (1983) and Branch (1977) find that the purchase of a security that has declined substantially by December has excess return in January. For example, Branch (1983) analyzed a trading rule that involved the purchase of a security that reached its annual low in the last week of trading in December. He found that these securities rose faster in the first four weeks of the new year than the market as a whole, with very little difference in risk. He obtained average returns 8% above the market for a four-week holding period. Reinganum (1983) found similar results.

For this to be a partial explanation of the January seasonal, it needs to be true that small stocks are an unusually high percentage of the stocks that are candidates for tax swapping. This is exactly what Reinganum (1983) finds. However, Reinganum argues it is not the full explanation because he still finds a January effect (although much smaller) for firms that show gains in the prior year. Securities that are being sold for tax-loss purposes are more likely to be at the bid in December. Thus the tax-selling hypothesis and microstructure explanation are likely to be partially measuring the same effect.

Several studies have provided evidence that is difficult to reconcile with the tax-selling hypothesis. Jones, Pearce, and Wilson (1987) study a period from 1821 to 1917, before the introduction of the income tax. They find a January effect that is not significantly different from the January effect found after the introduction of the income tax. Similarly, Japan and Belgium, which were found to have a January effect, do not have a capital gain tax. Furthermore, Australia has a non-December tax year so that if the extra returns were tax related, the effect should be present in a different month. However, there are excess returns in January for Australia.

[15]Maxwell (1998).

In an efficient market we should not observe a seasonal pattern. Investors observing high returns in January should start to purchase at the end of December to take advantage of the extra return. This adjustment of the pattern of investor purchases should cause the pattern to disappear. Furthermore, the explanations we have can explain only part of the extra return. Thus the January seasonal is difficult to reconcile with efficient markets.

Predicting Return from Past Return

In this section, we discuss the predictability of return from past return. In the first section we discuss short-term predictability. In the second section, long-run predictability is examined.

Short-term Predictability Tests of short-term predictability examine whether return in the prior period (usually a day or days) can predict today's return. The tests range from simple ways of using past return data to complex trading rules. We discuss a few representative tests from this voluminous literature on short-term price movements.

Correlation Tests Correlation tests are tests of a linear relationship between today's returns and past returns. A regression of the following form is estimated:

$$r_t = a + br_{t-1-T} + e_t \tag{17.1}$$

The term a measures the expected return, unrelated to previous return. Because most securities give a positive return, a should be positive. The term b measures the relationship between the previous return and today's return. If $T = 0$, then it is the relationship between today's return and yesterday's return. If $T = 1$, it is the relationship between today's return and the return two periods previously; e_t is a random number and incorporates the variability of the return not related to the previous return.[16]

In the process of estimating Equation (17.1), the researcher obtains the correlation coefficient between r_t and r_{t-1-T}. The square of the correlation coefficient is the fraction of the variation of today's return explained by the return shown on the right-hand side of the equation. For example, a correlation coefficient of 0.5 means that $(0.5)^2 = 0.25$ or 25% of the variation of the term on the left-hand side of the equation is explained by the term on the right-hand side.

Table 17.2 reports the results of one study examining the correlation between today's return and return in prior periods (both continually compounded). The first column is a test of the relationship between today's return and yesterday's return. The second column is a test of the relationship between today's return and the return two days prior. As discussed earlier, the square of the number in the table is a measure of how much of the variation in return the equation explains. For example, the largest number (in absolute magnitude) in the first column is -0.123, associated with Goodyear. This implies that relating yesterday's return to today's return explains $(-0.123)^2$ or 1.51% of the variation in today's

[16]Return has been defined both as change in price plus dividends divided by the prior period's price and as the log of the ratio of the price plus dividends divided by the prior period's price. The latter is the continuously compounded rate of return. In addition, some researchers have used change in price on both sides of the equation. It has been shown that for correlation tests it makes little difference which is used (see Granger, 1975). For example, if a test utilizing price changes shows no relationship, then a test utilizing log price relatives would also show no relationship. Equation (17.1) is clearly a linear equation. In any test, b could be no different from zero, suggesting no relationship between the previous price change and the next price change, and yet there may be a nonlinear relationship between successive price changes. For example, $P_t - P_{t-1}$ might be related to complex combinations of $(P_{t-1} - P_{t-2})$ raised to various powers.

Table 17.2 Daily Correlation Coefficients (from Fama, 1970)

	Lag				
Stock	1	2	3	4	5
Allied Chemical	0.017	−0.042	0.007	−0.001	0.027
Alcoa	0.118[a]	0.038	−0.014	0.022	−0.022
American Can	−0.087[a]	−0.024	0.034	−0.065[a]	−0.017
AT&T	−0.039	−0.097[a]	0.000	0.026	0.005
American Tobacco	0.111[a]	−0.109[a]	−0.060[a]	−0.065[a]	0.007
Anaconda	0.067[a]	−0.061[a]	−0.047	−0.002	0.000
Bethlehem Steel	0.013	−0.065[a]	0.009	0.021	−0.053
Chrysler	0.012	−0.066[a]	−0.016	−0.007	−0.015
Du Pont	0.013	−0.033	0.060[a]	0.027	−0.002
Eastman Kodak	0.025	0.014	−0.031	0.005	−0.022
General Electric	0.011	−0.038	−0.021	0.031	−0.001
General Foods	0.061[a]	−0.003	0.045	0.002	−0.015
General Motors	−0.004	−0.056[a]	−0.037	−0.008	−0.038
Goodyear	−0.123[a]	0.017	−0.044	0.043	−0.002
International Harvester	−0.017	−0.029	−0.031	0.037	−0.052
International Nickel	0.096[a]	−0.033	−0.019	0.020	0.027
International Paper	0.046	−0.011	−0.058[a]	0.053[a]	0.049
Johns Manville	0.006	−0.038	−0.027	−0.023	−0.029
Owens Illinois	−0.021	−0.084[a]	−0.047	0.068[a]	0.086[a]
Procter & Gamble	0.099[a]	−0.009	−0.008	0.009	−0.015
Sears	0.097[a]	0.026	0.028	0.025	0.005
Standard Oil (Calif.)	0.025	−0.030	−0.051[a]	−0.025	−0.047
Standard Oil (N.J.)	0.008	−0.116[a]	0.016	0.014	−0.047
Swift & Co.	−0.004	−0.015	−0.010	0.012	0.057[a]
Texaco	0.094[a]	−0.049	−0.024	−0.018	−0.017
Union Carbide	0.107[a]	−0.012	0.040	0.046	−0.036
United Aircraft	0.014	−0.033	−0.022	−0.047	−0.067[a]
U.S. Steel	0.040	−0.074[a]	0.014	0.011	−0.012
Westinghouse	−0.027	−0.022	−0.036	−0.003	0.000
Woolworth	0.028	−0.016	0.015	0.014	0.007

[a]Coefficient is twice its computed standard error.

return. This is extremely small. The negative number implies that today's return is affected negatively by yesterday's return.

Despite the small size of the numbers, looking at the table might provide some evidence in favor of a weak relationship between returns over time. Twenty-two of the 30 numbers are positive, which is fairly high if there is no relationship. Furthermore, 11 of the numbers are significantly larger than would be expected by chance (although 2 of these are negative). Once again, this is more than one would expect. However, lest one get too excited by the relationship, the average absolute value of column 1 is 0.026. This implies that 0.067% of the variation in today's return is explained by yesterday's return.

Earlier we noted that a was the expected return in Equation (17.1). Investigators using correlation tests are, in essence, fitting Equation (17.1) to a body of data. The estimate of expected return arrived at for a security is the average unexplained by past return. This is

very close to the average historical return. Different results might be obtained if the term *a* were set equal to different estimates of expected return. In the next section we see that in semistrong tests of the EMH, expected return is usually obtained by the single-index model discussed in Chapter 7. It is possible that there may be a different correlation in the return series when average return is defined using some other model such as the Single Index Model.

Although serial correlation effects are not strong, they have remained in the data since Fama's early study. Explanations for the phenomena have focused on microstructure effects. For example, Campbell, Grossman, and Wang (1993) show that variation in trading volume explains some serial correlation.[17]

Another version of the random walk test is based on variance ratios, that is, the relationship of the volatility of a single-period return to the volatility of a multiperiod return. Let $R_{t,\,t+T}$ equal the return of a stock over T periods from time t to time $t+k$:

$$R_{t,T} = R_t + R_{t+1} + \cdots R_T \tag{17.1}$$

If returns are serially independent of each other, then their variances are additive:

$$\sigma^2_{R_{t,t+T}} = \sigma^2_{R_t} + \sigma^2_{t,t+1} + \cdots \sigma^2_{R_{t,t+T}} \tag{17.2}$$

If the variances of the individual-period returns are equal to each other, then the variance of the multiple-period returns is proportional to the square root of the number of periods:

$$\sigma_{R_{t,t+T}} = \sqrt{T}\ \sigma_{R_i} \tag{17.3}$$

Regnault was the first to propose this $\sqrt{T}$ relationship with daily French government bond prices. Lo and MacKinley (1988) use this same relationship to test the random walk hypothesis for weekly returns to portfolios of U.S. stocks. They found that small stocks in particular violated the pure random walk model.[18] While the variance ratio test is a natural implication of random walk models, it may not be the most robust method of testing serial independence. Deo and Richardson (2003) show that this test is weak and potentially biased.[19] Furthermore, as Brown (2011) observes, it may lead to false rejections of the EMH. The random walk hypothesis, unlike the EMH, makes the strong assumption that successive changes in the logarithm of price are distributed with fixed and finite variance. Changes in market risk of the kind made evident in the market crisis of 2007–2008 violate the assumptions behind the random walk hypothesis and would lead to an apparent rejection of the model based on data taken from periods when the market variance is nonstationary.

Correlation for Portfolios of Securities There is evidence of somewhat higher correlation between past return and future return for portfolios of stocks compared to individual stocks. Lo and MacKinlay (1988) and Conrad and Kaul (1988) put together portfolios that are grouped by size (number of shares times price per share). They find that this week's return is related to the prior week's return and that this relationship is stronger for portfolios of small stocks. The weekly correlation coefficients from Conrad and Kaul for the largest portfolios is 0.09, so that 0.81% of this week's return is explained by the prior week's. However, the correlation coefficient for the smallest portfolio is 0.3, implying that 9% of this week's return can be explained by return in the prior week. The results suggest that because of the variance reduction of diversification, correlation of weekly returns is higher for

[17]Campbell, Grossman, and Wang (1993); Chordia and Swaminathan (2002).
[18]Lo and MacKinlay (1988).
[19]Deo and Richardson (2003).

portfolios than individual stocks. However, one must be somewhat cautious in interpreting these results. Portfolios will show correlation between past return and future return because some securities do not trade continually, and important information might be reflected in the securities at different times. Thus a major release of market information might affect securities in different weeks, causing returns to be correlated not because past returns predict future returns but because of infrequent trading. This latter hypothesis is consistent with the correlation of portfolios being greater for small portfolios than large portfolios.

Correlation over Long-Run Horizons

Fama and French (1988) and Poterba and Summers (1988) have examined the correlation in returns computed over longer periods. Fama and French (1988) find, using data from 1926–1985, that the correlation between this period's returns and return in the prior period is -0.25 for three-year periods to -0.40 for five-year periods. Poterba and Summers find similar results using a somewhat different methodology.

Fama and French argue that these results should not be given a lot of weight because both their procedures and those of Poterba and Summers have very little statistical power (could easily result from chance), and because the correlation is much smaller and insignificant after 1940. Furthermore, Fama (1991) argues the results could be due to a combination of a changing expected return and expected return reverting to its mean over time.

The variance ratio discussed earlier is particularly important in tests of mean-reversion in longer-horizon returns. Poterba and Summers (1988) use the variance ratio methodology to document an apparent violation of the random walk model in long-horizon returns to the U.S. stock market. They find evidence consistent with mean-reversion. Notice from Equation (17.2) that the variances sum only because of independence. We know from portfolio theory that the variance of the sum of N negatively correlated random variables will have lower variance than times the average variance. Fama and French (1988) measure the correlation in returns computed over multiyear horizons using data from 1926–1985 and found that the correlation between this period's returns and return in the prior period is -0.25 for three-year periods to -0.40 for five-year periods. This finding had two important implications. First, it suggested that the stock market did not follow a random walk and may not be efficient. More importantly for investment management and long-term planning, they found that longer-horizon returns were less volatile than predicted by standard models. In the short run the market appears to overshoot and in the long run to revert back to prior levels. Barberis (2002) showed how a long-term investor could take advantage of this pattern. One challenge to scholars testing mean reversion at longer horizons is that the number of independent observations available to test decreases with the desired investment horizon. Fama and French had only 60 years of retums to study, leaving only 12 independent five-year observations. Using overlapping five-year intervals created problems of statistical inference. One solution to the limited data problem was to obtain longer time series. Goetzmann (1993) constructed an index for the London Stock Exchange from 1695 and found evidence of mean reversion around long-term trends. Goetzmann, Ibbotson, and Peng (2001) constructed an index for the NYSE from 1815 and found that the period prior to 1870 exhibited strong evidence of mean reversion at long horizons, however, the correlations change and become weaker in the post-1870 period. The implication of these longer-term studies of mean-reversion is that, while there is some evidence of predictability due to mean-reversion, the verification periods are long and thus unlikely to be amenable to correction by arbitrage. They might, however, be useful for long-term planning. If the stock market for long-term investors is slightly less risky than its short-term dynamics suggest, then this is an argument for adjusting inputs to the optimization process and perhaps holding more equities for the long term.

Runs Tests Most of the tests of the usefulness of past return in predicting future return utilize correlation coefficients to examine efficiency. The correlation coefficient tends to be heavily influenced by extreme observations. Thus results can be due to one or two unusual observations. An alternative analysis, which eliminates the effect of extremely large observations, is to examine the sign of the price change. Designate a price increase by + and a price decrease by −. Then, if price changes were positively related, it would be more likely that *a* + was followed by *a* + and *a* − by *a* − than to have a reversal in sign. This would mean that an investigator analyzing a sequence of correlated price changes would expect to find longer sequences of +s and −s than could be attributed to chance. A sequence of the same sign is called a *run*. Thus, + − − − + + + 0 has four runs, a run of one +, a run of three −s, followed by a run of three +s, followed by a run of one no change. If there were a positive relationship between price changes, there should be more long sequences of + and − than could be attributed to chance and fewer runs.

Many of the authors who examined correlation also examined runs. Table 17.3 is a typical example taken from Fama (1965). For one-day intervals, 760 runs were expected and

Table 17.3 Total Actual and Expected Numbers of Runs for 1-, 4-, 9-, and 16-Day Differencing Intervals (from Fama, 1965)

	Daily		4-Day		9-Day		16-Day	
Stock	Actual	Expected	Actual	Expected	Actual	Expected	Actual	Expected
Allied Chemical	683	713.4	160	162.1	71	71.3	39	38.6
Alcoa	601	670.7	151	153.7	61	66.9	41	39.0
American Can	730	755.5	169	172.4	71	73.2	48	43.9
AT&T	657	688.4	165	155.9	66	70.3	34	37.1
American Tobacco	700	747.4	178	172.5	69	72.9	41	40.6
Anaconda	635	680.1	166	160.4	68	66.0	36	37.8
Bethlehem Steel	709	719.7	163	159.3	80	71.8	41	42.2
Chrysler	927	932.1	223	221.6	100	96.9	54	53.5
Du Pont	672	694.7	160	161.9	78	71.8	43	39.4
Eastman Kodak	678	679.0	154	160.1	70	70.1	43	40.3
General Electric	918	956.3	225	224.7	101	96.9	51	51.8
General Foods	799	825.1	185	191.4	81	75.8	43	40.5
General Motors	832	868.3	202	205.2	83	85.8	44	46.8
Goodyear	681	672.0	151	157.6	60	65.2	36	36.3
International Harvester	720	713.2	159	164.2	84	72.6	40	37.8
International Nickel	704	712.6	163	164.0	68	70.5	34	37.6
International Paper	762	826.0	190	193.9	80	82.8	51	46.9
Johns Manville	685	699.1	173	160.0	64	69.4	39	40.4
Owens Illinois	713	743.3	171	168.6	69	73.3	36	39.2
Procter & Gamble	826	858.9	180	190.6	66	81.2	40	42.9
Sears	700	748.1	167	172.8	66	70.6	40	34.8
Standard Oil (Calif.)	972	979.0	237	228.4	97	98.6	59	54.3
Standard Oil (N.J.)	688	704.0	159	159.2	69	68.7	29	37.0
Swift & Co.	878	877.6	209	197.2	85	83.8	50	47.8
Texaco	600	654.2	143	155.2	57	63.4	29	35.6
Union Carbide	595	620.9	142	150.5	67	66.7	36	35.1
United Aircraft	661	699.3	172	161.4	77	68.2	45	39.5
U.S. Steel	651	662.0	162	158.3	65	70.3	37	41.2
Westinghouse	829	825.5	198	193.3	87	84.4	41	45.8
Woolworth	847	868.4	193	198.9	78	80.9	48	47.7
Averages	735.1	759.8	175.7	175.8	74.6	75.3	41.6	41.7

735 were obtained. Thus there were fewer runs than were expected, which is evidence of a small positive relationship between successive returns. The results for longer intervals are very striking. The actual number of runs in each case was almost exactly equal to the expected number.

In summary, correlation and runs tests seem to show some small positive relationship between today's return and yesterday's return, but on average it is very small and frequently negative for individual securities.

Some correlations could be observed and the market still be efficient. An investor must incur transaction costs to trade securities. Thus, if the correlation is very low, transaction costs should more than eliminate any potential profits from attempting to take advantage of correlated series. In fact, in an efficient market, transaction costs would set an upper limit to the amount of correlation. One indication that markets are efficient would be if we observed higher correlation in markets with higher transaction costs. This is exactly what Jennergren and Korsvold (1975) found when they examined the higher transaction costs of Norwegian stocks.[20]

Although this is an indication that the correlation is insufficient to cover transaction costs, more direct tests are necessary. It is to these tests and tests of more complicated ways of using past return that we now turn.

Filter Rules We have discussed tests of whether returns are linearly related to past returns. Even in the absence of such regular and simple patterns, it is possible that complex patterns exist that allow excess profits to be made. The simplest way to test for the existence of more complex patterns is to formulate a trading rule appropriate for a particular pattern of returns and see what would have happened if one had actually traded on these rules.

One price pattern that has frequently been hypothesized for price movements is depicted in Figure 17.1. The argument behind this figure proceeds as follows. As long as no new information enters the market, the price fluctuates randomly within the two barriers around the "fair" price. If the actual price differs too much from the "fair" price, then "professionals" will step in and purchase or sell the security. This will keep the security price within the security price barriers. However, if new information comes into the market, then

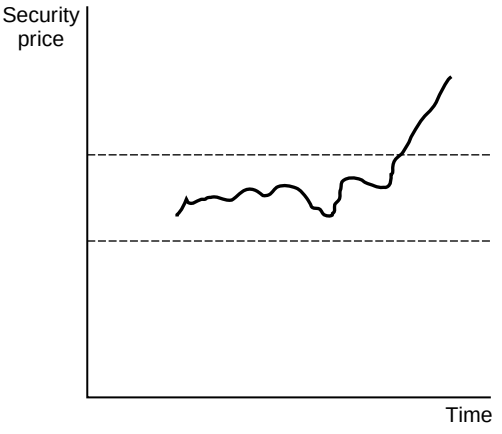


Figure 17.1 Security price and time.

[20]Jennergren and Korsvold (1975) found 338.2 runs per stock over a period when uncorrelated returns would have led to 394.6.

a new equilibrium price will be determined. If the news is very favorable, then the price should move up to a new equilibrium, well above the old price. Investors will know that this is occurring when the price breaks through the old barriers. If investors purchase at this point, they will benefit from the price increase to the new equilibrium level. Similarly, if bad news concerning the company is forthcoming, the stock will drop to a new equilibrium level. If investors sell the stock as it breaks the lower barrier, they will avoid much of the decline. If they sell the stock short as it breaks through the barrier, they will benefit from the decline. This argument is intuitively appealing; it is closely analogous to the idea of control charts and is put forth as an appropriate investment strategy by many who believe price series can be used to make superior profits. The strategy is called a *filter rule*. The filter rule is usually stated in the following way: Purchase the stock when it rises by *X*% from the previous low and hold it until it declines by *Y*% from the subsequent high. At this point, sell the stock short or hold cash.

Filter rules are a timing strategy. They show investors when they should be long in a security and when they should sell it short. The alternative to timing is to buy and hold the security. Thus filter rules are analyzed by comparing them to a buy-and-hold strategy.[21] The classic tests of filter rules were performed by Fama and Blume (1966). They found that small filters, for example, 0.5% were profitable before transaction costs but not after transaction costs. The profitability of these very small filter rules is consistent with a slight positive correlation of security price changes and with the evidence discussed earlier.

Jennergren and Korsvold (1975) found some of the highest correlation coefficients of any investigators when they examined the lightly traded Norwegian and Swedish stocks. The relatively high correlations suggest that these securities are prime candidates for profitable filter rules. Jennergren examined filter rules for these securities. Norwegian and Swedish stocks cannot be sold short so that the alternative to holding securities long was to invest in a savings account. Some of the filter rules outperformed a buy-and-hold strategy. When taxes and transaction costs were considered, however, only the king (the only tax-exempt investor) had any prospects of making a profit.

We have examined one type of filter rule that purports to aid in timing decisions. We could test other types that suggest trades on the basis of alternative price patterns. Indeed, technical analysts are fond of talking about such things as head-and-shoulder patterns and other esoteric perceived price phenomena. But there is no evidence that trading on the basis of any of these patterns can lead to an excess profit.

Returns and Firm Characteristics

In this section we examine firm characteristics and returns. In particular we examine what characteristics of firms are associated with excess returns. It has been found that a number of firm characteristics such as size, market value divided by book value, and earnings divided by price are related to excess return.

The relationship between firm characteristics and excess returns is a difficult set of empirical findings to reconcile with the concept of efficient markets. Indeed these are often referred to as market anomalies, because in an efficient market it should not be possible to earn an excess return on the basis of observable firm characteristics.

[21]A number of tests of filter rules have analyzed returns during periods of market decline. During these periods, any rule that randomly caused the investor to sell a security and hold cash or go short should, on average, outperform a buy-and-hold strategy, at least before transaction costs are considered. The filter rule is purported to be a rule that utilizes past price behavior to lead to superior timing. It is important (if the rule is tested during periods of price decline) to determine that the rule outperforms a rule that randomly causes an investor to sell the security.

There are five possible explanations for the existence of a relationship between firm characteristics and excess returns.

The first explanation is that the relationship observed is not real. With hundreds of researchers examining the same data for patterns, some relationship between firm variables and returns will be found. Furthermore, the conventional statistical test utilized to examine the statistical significance of the relationship they found is inappropriate, because they test the likelihood of one study finding a relationship, not one study out of hundreds of studies. Thus the tests that find a significant statistical relationship overstate the significance.

The second explanation is that these firm characteristics serve as a proxy for an omitted risk variable and that once this variable is taken into account, the relationship between firm characteristics and excess return disappears. For example, small firms have excess returns when measuring expected return using the CAPM. However, some researchers argue that small firms have lower probability of survival and that "survival probability" isn't adequately measured by beta. Furthermore, once this risk variable is taken into account, the excess returns associated with size disappear.

The third explanation is that the CAPM is a reasonable model of expected returns but has been misestimated, causing apparent large returns when none exist. For example, assume betas are systematically underestimated for small firms; then the estimate of expected returns for small firms would be too low and they would appear to have excess return when none would exist if betas were estimated properly.

A fourth explanation of why the phenomenon can continue to exist in an efficient market but not why it occurs in the first place is that trading costs eliminate the profitability of any trading rules designed to exploit the strategy.

Finally, markets may simply be inefficient.

The "Size Effect" Banz (1981) published one of the earliest and most often quoted empirical articles on the size effect. Employing a methodology similar to that used by Fama and MacBeth (1973) (see Chapter 15), Banz documented that excess returns (alphas) would have been earned over the period 1936–1977 by holding small firms. The striking aspect of Banz's analysis is that the size effect appeared to be important in terms of both statistical significance and empirical relevancy. The size term had roughly the same statistical significance in explaining returns as did beta. Furthermore, the differential returns from buying very small firms versus very large firms were 19.8% per year. Other points should be mentioned. The real payoff from holding small stocks came from holding the smallest 20% of the firms in Banz's sample of NYSE firms. The differential between other quintiles was quite small. Second, although on average the return from holding the smallest firms was large and statistically significant, there were periods of time where large firms outperformed small firms.

Subsequent to Banz's study, it has been documented that a substantial part of the size effect occurs in January. For example, Keim (1983) reports that the difference in the returns in January due to size are about half of the annual difference. Thus the size effect and January effect are strongly related.

The size effect was the first of the firm variables that was shown to be related to excess return; there has been extensive research into possible explanations. One research avenue has been to hypothesize that the CAPM was inappropriately measured, causing apparent excess returns. The argument is that the betas estimated for small firms were too low. If beta is too low, then the estimate of expected return using the CAPM is too low and the difference between actual return and expected return would be positive even if it was zero when expected return was correctly estimated. Two reasons have been offered for why estimated betas are too low for small firms. Roll (1970) and Reinganum (1981) have shown that the

beta for small firms will be biased downward because they trade less often than large firms, and nonsynchronous trading leads to an underestimate of beta. Christie and Hertzel (1981) present a second reason why beta might be downward biased. Beta is measured using historical returns. Firms that become small have changed their economic characteristics; these changes mean they are riskier, and beta measured over a prior period doesn't capture this increased risk. These factors could partially explain the relationship of excess return to size.

A second approach to explaining the small firm effect is to argue that expected return was miscalculated because the CAPM or zero beta CAPM are inappropriate models for measuring expected return. Perhaps a multifactor model better explains expected returns, and when these models are used to measure expected return, the size effect disappears. An example of this research is Chan, Chen, and Hsieh (1985). They use the APT model of Chen, Roll, and Ross to measure expected return on 20 portfolios formed on the basis of size. They find that the difference in return between the smallest portfolios and the largest portfolio was 1.5% per year. In contrast, using the standard CAPM resulted in a difference in return of 11.5% per year. Thus they conclude that the size effect disappears when a more appropriate model of expected returns is used. The additional variable in their APT model that explains most of the variation in return between portfolios of different size is the difference in return between high-risk corporate bonds and government bonds. In a later paper Chan and Chen (1991) argue that the reason small firms are riskier is that they have low production efficiency and high leverage, and are in their terms "marginal firms" with lower probability of surviving economic hard times. They point out that size is serving as a proxy for this more fundamental risk.

Another reason why the CAPM may misestimate expected return was studied by Amihud and Mendleson (1991). They reason that investors should demand a higher expected return for less liquid stocks because trading them involves higher transaction costs. Empirically small stocks have higher bid–ask spreads, and the price impact of larger purchases would be considerable for small stocks. Thus they show the small stock effect is in part compensation for illiquidity.

Finally, a number of researchers have argued that transaction costs are very high in small stocks, so that markets are still efficient with substantial excess returns on small stocks. First, Roll (1983) and Blume and Stambaugh (1983) have estimated that the magnitude of the small firm effect is cut in half if small stock portfolios are reformed annually rather than rebalanced daily, as assumed by a number of authors. If the reader wonders why not simply buy small firms and rebalance daily, the answer is that large transaction costs would be incurred. Second, a number of authors have estimated transaction costs for small stocks and then argued that the excess return is eliminated or at least reduced if realistic transaction costs are taken into account.[22]

Market to Book Fama and French (1988), Lakonishok, Shleifer, and Vishny (1993), and Chan, Hamao, and Lakonishok (1991) have all examined the relationship between market to book and excess return or return.

For example, Lakonishok, Shleifer, and Vishny (1993) examine returns on portfolios of stocks bought on the basis of a stock's book to market value. To control for size effects, they first classify stocks into five size categories. Within each of the five size categories they classify stocks into 10 equal-size groups on the basis of market to book value. The average difference in return between the high-book-to-market firms and the low-book-to-market firms is 7.8% per year. They attempt to examine whether this difference could be explained by risk.

[22]There is a counterargument. Small stock index funds are able to match the small stock index. They do this in part by utilizing trading strategies that reduce transaction costs. Thus the estimates of transaction costs that have been presented may not be a realistic estimate for portfolio managers.

The normal procedure would be to use one of the equilibrium models of Chapters 13–16. However, they take a different and a very interesting approach. They separate out good market periods and bad market periods. They argue that if a stock is less risky, it is because it gives its good outcomes when it is needed most, namely, in bad markets. They find that low-market-to-book stocks do not give a higher return when markets are poor, and thus argue that the higher return on high-market-to-book firms is not compensation for risk.

Earnings Price Basu (1977) has shown that when expected returns are measured by the CAPM model, excess returns (return minus expected return) are positively related to the firm's earnings/price (E/P) ratio.

There has been much less work on the E/P effect than the size effect. Reinganum (1981) presents empirical evidence that the E/P effect is highly correlated with the size effect. Fama and French (1989) argue that once size and market to book are accounted for, the E/P effect disappears.

Chan, Hamao, and Lakonishok (1991) get similar results. Thus most researchers have seen the E/P relationship as a proxy for other effects.[23]

Predicting Long-Run Returns from Firm and Market Characteristics

Long-run returns of bonds and common stocks seem to be predictable using past variables related to the general level of the stock market and the term and risk structure of interest rates. For example, five-year returns on the Standard and Poor's (S&P) index might be regressed on the dividend price ratio and the difference in yield on corporate bonds compared to government bonds. Because the dividend price ratio and the yield difference is known at the beginning of the period, this relationship could be used to predict returns in subsequent periods. Some authors interpret evidence of predictability as showing that expected return changes over time and that these changes can be predicted. Other authors view this evidence as an indication of inefficiency in the stock and bond markets.

The variables that have been used to predict return include the following:

LEVEL OF MARKET VARIABLES

1. Dividends on S&P index/price of S&P index
2. Earnings of S&P index/price of S&P index
3. Current S&P index/long-run average of S&P index

INTEREST RATE VARIABLES

1. Term premium (yield on long-term bonds minus yield on short-term bonds)
2. Risk premium (yield on low-rated debt minus yield on high-rated debt)

The proportion of long-term return that can be explained by these variables is quite high.[24] Fama and French (1966) report that 25% of the returns on a value and equally weighted market index over two to four years can be explained by past dividends/price. Furthermore, the sign is positive: a high dividend over price (low level of price) implies high returns. Similarly, Campbell and Shiller (1988) find that earnings/price where earnings are averaged over 30 years can explain more than 57% of the yearly returns on a market index.

[23]Basu (1983) was able to find an earnings price effect even after adjusting for size. However, he did not simultaneously control for market/book. This is probably because the market/book effect has only been documented recently.
[24]Goetzmann (1993) and Goetzmann, Ibbotson, and Peng (2001).

In a later article Fama and French (1991) find that dividend/price plus the term and risk premiums explain a significant proportion of the returns not only for the aggregate stock market index but also for an index of small stock returns and indexes of high- and low-grade bonds. Furthermore, they find that the effect of dividend over price and risk spread bear a logical relationship to the return on the different instruments. For example, an increase in dividend/price predicts a greater increase in return for small stocks compared to large stocks, stocks compared to bonds, and low-grade debt compared to high-grade debt.

Finally, Harvey (1991) finds that the S&P dividend/price and the U.S. term structure variables predict long-term returns on portfolios of foreign stocks.

Although it has become a standard assumption in financial research that long horizon stock returns are predictable by a set of forecasting variables, this has been qualified to the same extent by other research. Richardson (1993) and others pointed out that the statistical tools used in long-horizon prediction studies have problems with inference. Goetzmann and Jorion (1993, 1995) show that the forecasting power of dividend yields is compromised by statistical biases. Other studies demonstrating the weakness of the long-horizon predictability evidence include Welch and Goyal (2008).[25] Welch and Goyal find no convincing evidence that fundamental ratios such as the dividend yield and the earnings price ratio could have been used out-of-sample to beat the market. Ang and Bekaert (2007), using longer data series, international markets, and a robust, nonlinear specification, find that "at long horizons, excess return predictability by the dividend yield is not statistically significant, not robust across countries, and not robust across different sample periods."[26] Thus, although a number of studies suggest that firm and market characteristics could have been used to time the market, applying these results to forward-looking investment planning should be done cautiously.

With respect to the basic question of testing the EMH, it is important to note that return predictability does not necessarily violate efficiency: it could simply reflect changes over time in expected returns or the risk premiums associated with priced factors. In fact, this is part of a broader problem, namely, that no firm conclusions about efficiency are possible without the "correct" model of expected returns. Unless two return series are perfectly correlated, a difference in their average returns could be due to a risk factor that has not been properly accounted for in the model of expected returns. This "joint hypothesis" problem affects analyses of efficiency in all asset classes and, in particular, in equities and fixed income. Should an investor seek to time his or her exposure to the equity premium through long-horizon forecasting models? Although it is useful to recognize the potential value to be added by this approach, evidence of its efficacy is mixed at best.

ANNOUNCEMENT AND PRICE RETURN

The greatest amount of research in finance has been devoted to the effect of an announcement on share price. These studies are known as "event studies." Initially event studies were undertaken to examine whether markets were efficient, in particular, how fast the information was incorporated in share price.

For example, when a firm announces earnings will be much larger than expected, will this news be reflected in share price the same day or over the next week? Dozens of studies confirmed that share prices reacted rapidly to announcements, and in expected ways, where the direction of the price change and the likely impact were clear. Consequently,

[25]Welch and Goyal (2008).
[26]Ang and Bekaert (2007).

many authors accept that information is rapidly incorporated into share price and use event studies to determine what information is reflected in price and, if its impact is unclear, to determine whether the announcement is good or bad news.

METHODOLOGY OF EVENT STUDIES

The methodology of event studies is fairly standard and proceeds as follows:

1. **Collect a sample of firms that had a surprise announcement (the event).** What causes prices to change is an announcement that is a surprise to investors. For many studies, any announcement, such as an announcement of a merger, can be treated as a surprise. For other studies, such as the impact of earnings announcements, it is more complicated. For these studies, it is necessary to define a surprise. This is normally done by comparing announcements to what was expected as reflected in the average estimate of professional analysts. A number of services provide these data. To form a sample of surprises, one first separates out a group of firms where the announcement is significantly different from what is being forecast. Because positive and negative surprises would affect price differently, this group is further separated into two groups, one for positive and one for negative earnings surprises.
2. **Determine the precise day of the announcement and designate this day as zero.** Most current studies use daily data, whereas the original studies used monthly data. The use of monthly data made measurement much more difficult because there are many surprises in a month besides the announcement effect being studied. Thus, for measuring market efficiency, it is important to measure the impact of the announcement using the smallest feasible intervals. A number of recent studies have used intraday data.
3. **Define the period to be studied.** If we studied 60 days around the event, then we would designate $-30, -29, -28, \ldots, -1$ as the 30 days prior to the event, 0 as the event day, and $+1, +2, +3, \ldots, +30$ as the 30 days after the event.
4. **For each of the firms in the sample, compute the return on each of the days being studied.** In the example, this is 61 days (30 before the event plus the event day plus the 30 days after the event).
5. **Compute the "abnormal" return for each of the days being studied for each firm in the sample.** Abnormal return is actual return less the expected return. Different authors use different models for expected return. Any of the equilibrium models discussed in Chapters 13, 14, and 16 could be used to define expected return. Other authors use the market model of Chapter 7. Finally, a number of studies simply use the return on a market index as the expected return.[27]
6. **Compute for each day in the event period the average abnormal return for all the firms in the sample.** When this is done, we can examine the data in a figure such as Figure 17.2. We normally look at the average effect of the announcement rather than examine each firm separately, because other events are occurring, and averaging across all firms should minimize the effect of these other events, thereby allowing a better examination of the event under study. However, for studies where the magnitude of the announcement should vary across firms (such as earnings surprises), it may be useful to examine individual firm behavior as well.

[27]See Brown and Warner (1985) for a comparison of these techniques. One must be sure that the sample does not have special characteristics, such as small size, that have been shown to produce abnormal returns.

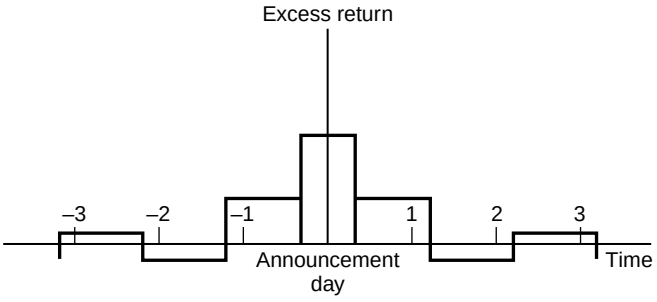


Figure 17.2 Excess return around announcement day.

7. **Often the individual day's abnormal return is added together to compute the cumulative abnormal return from the beginning of the period.** In this case, for a 61-day period (30 before the event day, and 30 days after) the entry for −20 would be the sum of the daily average abnormal returns for days −30 to −20 and the entry for −10 would be the sum of the average daily abnormal returns for −30 to −10. Using the data for average daily abnormal returns shown in Figure 17.2, this produces a chart such as that shown in Figure 17.3. Notice that Figure 17.2 has a large positive abnormal return shown on day zero and nothing but randomness on other days. However, in Figure 17.3, which is the cumulative abnormal return, the positive abnormal return on day 0 persists because it is part of the cumulative returns on days +1 through +30.

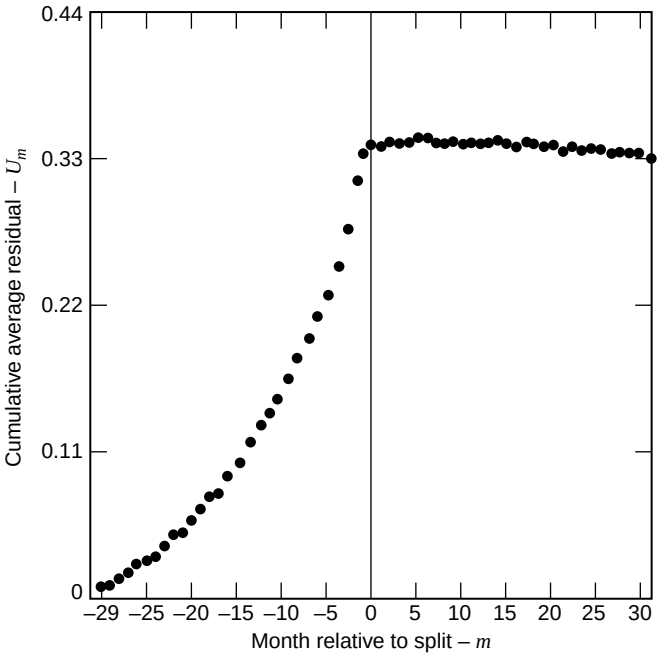


Figure 17.3 Cumulative excess return around split rate.

Figures 17.2 and 17.3 are the pattern of abnormal returns you would expect to find if markets are semistrong-form efficient. Thus, on the day of the announcement, you would expect an abnormal return, but not on other days. However, normally some abnormal return is found on the days surrounding the announcement. Abnormal return after the announcement day is either due to information taking time to be reflected in share price or the announcement taking place so late on day 0, possibly even after the markets close, that its effect can only be reflected in trades and prices on the day following the announcement. Abnormal returns prior to the announcement day can come from three sources. First, the fact that an important announcement will take place is often released to the public prior to the announcement, and the news release that an announcement will take place and the way the release is handled may convey information. Thus a message conveyed to analysts and the financial press that there will be an important announcement at a luxury hotel with drinks and hors d'oeuvres afterward may convey information that there will be a welcome surprise. In an efficient market this should be reflected in price before the announcement takes place.

Second, if the announcement is at the discretion of the firm, it may be partially caused by prior abnormal returns, and an event study of this announcement will show prior abnormal returns. For example, firms split their stock generally after a substantial price rise. Event studies of stock splits will find abnormal returns prior to the announcement because firms with abnormal returns are more likely to split their shares. Third, abnormal returns prior to the announcement day could reflect leakage of the information by those with access to it.

8. **Examine and discuss the results.** Having performed the analysis, the results are examined and conclusions drawn.

Results of Some Event Studies

We will not review all types of event studies in this chapter. Rather, we will concentrate on issues that are especially important for investment strategy. In particular, in this section we will examine the pattern of abnormal returns around the announcement day and whether there is a long-term abnormal return after the announcement (postannouncement drift). These questions are concerned with whether an investor can make short-term profits by buying on the announcements or make long-term abnormal profits by buying on the announcements and holding (or short selling if the drift is downward) over the longer period of time. Both strategies provide evidence on market efficiency.

The interpretation of abnormal returns earned around the announcement day is fairly noncontroversial. If annual market returns are 10%, then daily market returns are about 0.04%. Because most studies find abnormal returns of several percent at the time of the announcement, any way of measuring expected returns will show about the same results, unless announcements are clustered on days of extreme market movements. Thus, how expected returns are calculated is not important in interpreting results on event days. In the following pages we will discuss three typical studies.

A number of studies have examined whether markets are efficient with respect to the announcement of the purchase or sale of securities (see Kraus and Stoll, 1972; Grier and Albin, 1973; Dodd and Ruback, 1977). In general, these studies find that markets are efficient. One of the more interesting studies of this type was by Firth (1975). He examined the efficiency with respect to an announcement that an individual or firm had acquired 10% of a firm. In the United Kingdom (which Firth analyzed) as well as in the United States, ownership of more than some percentage must be made public. Firth examined the market efficiency with

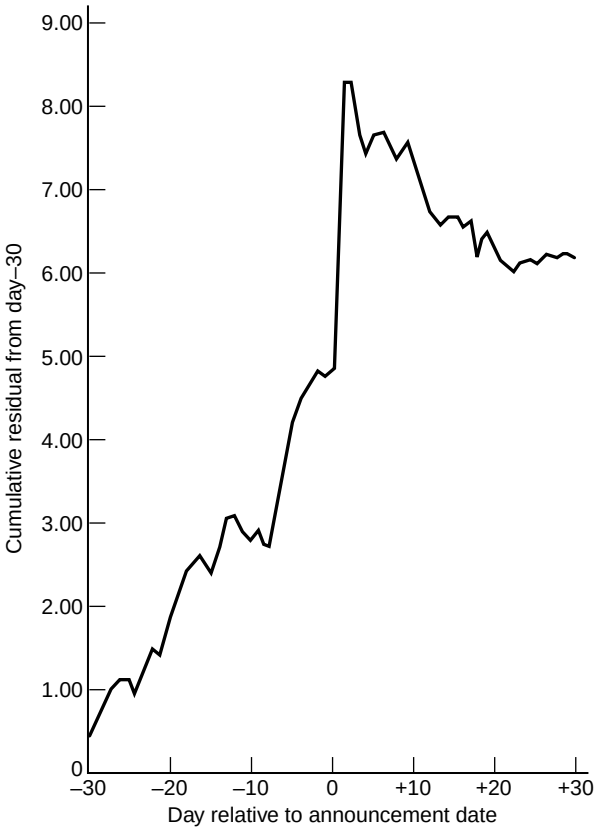


Figure 17.4 Cumulative excess return around announcement date.

respect to these announcements. One would expect that the purchase of a substantial percentage of a company might be an indication of a takeover or merger attempt, and Firth showed that this is an appropriate expectation. Empirical evidence indicates that mergers and takeovers normally involve premiums being paid to the stockholders of the company being taken over. Thus, the announcement of someone taking a large position in a security should be an indication of favorable prospects. Firth uses the single-index model to calculate expected return.

Figure 17.4 shows the cumulative excess returns from 30 days prior to the announcement. The cumulative excess return through the first day after the announcement is, in general, increasing. An investor with inside information that someone was accumulating a large block could make excess profits possibly larger than transaction costs. There is a substantial increase in cumulative excess returns on the day of the announcement. However, Firth shows the bulk of this increase occurs between the last trade before the announcement and the next trade. Thus an investor without prior information about the announcement could not benefit from the price increase. From the first trade after the announcement until 30 days after the announcement, there is a slight decline in the cumulative excess return. In general, this evidence is consistent with market efficiency.

Another example of semistrong-form tests of efficiency was performed by Davies and Canes (1978). They analyzed whether analysts' information could be used to earn excess returns or if it was already incorporated into share price. An enormous amount of information

is sold to investors, including stock recommendations as well as detailed information on individual securities. One would expect that recommendations that are purchased contain sufficient information to justify their cost. Davies and Canes (1978) analyze this by examining the usefulness of the "Heard on the Street" column in the *Wall Street Journal.* This column usually consists of a number of opinions on different stocks. The publication of the analysts' opinions in the *Wall Street Journal* usually occurs one or two weeks after the opinion was circulated to the firms' clients. However, the *Wall Street Journal* is usually the first large-scale dissemination of the opinions of several analysts.

The method of analysis was very similar to that discussed previously. The market model was used to estimate the relationship between each security's return and the market.[28] This equation was then used to estimate the expected return on each day given the actual level of the market. The difference between actual return and expected return was then tabulated. Figure 17.5 shows the results. As can be seen by examining the figure, the publication of the information seems to have an impact on returns. Davies and Canes tested to see if the large differences were statistically significant and found that they were.

The excess returns on the printing in the *Wall Street Journal* indicate that the column contains information that investors had not received directly from the analysts or that the material in the *Wall Street Journal* conveys information possibly by certifying the analysts' recommendations.

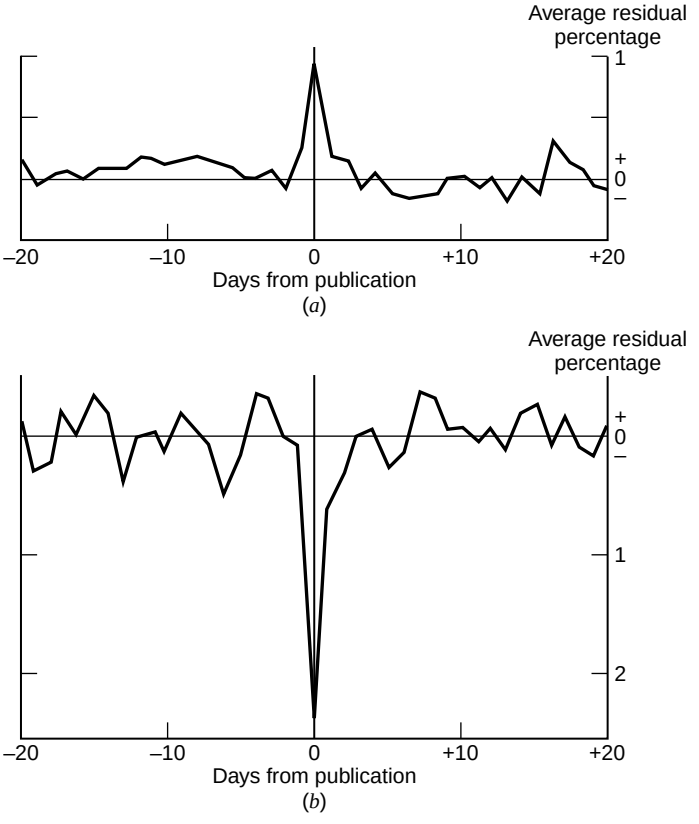


Figure 17.5 Excess return around publication date.

[28]The market model is the single index of Chapter 7 without the assumption of uncorrelated residuals.

As a final example of announcements and market efficiency tests, consider the examination of dividend announcements by Pettit (1972), Watts (1973), Charest (1978), Aharony and Swary (1980), and Agrawal and Mullins (1983). Two aspects of these studies are different from those discussed previously. First, they must carefully define the event relative to expectations. What should affect security prices is surprises, not events that are anticipated. It is reasonable to assume that an announcement of a stock split, or the acquisition by one investor of a large position in a security, is a surprise. However, changes in dividends may well be anticipated. It has been shown that firms tend to follow a stable dividend policy. Thus, when earnings increase, the firm may have a policy of increasing dividends. This implies that a dividend increase may have been anticipated.

To determine whether the dividend is good news (above anticipations), bad news (below anticipations), or no news (anticipated), each of the authors employs a model of dividend policy. For example, Watts relates changes in dividends to the level of previous dividends and earnings. Firms are then dichotomized into two groups: those firms whose dividends are above those predicted using the model and those that are below. Examining the excess return for these two groups allows one to examine the effect of unanticipated dividend changes.

The second difference in these studies is the need to disentangle the dividend changes from other effects. For example, stock splits and dividend increases often occur simultaneously, but in some cases they do not. Furthermore, dividend announcements almost always occur simultaneously with earnings announcements, and it is important to deal with contemporaneous earnings surprises. Pettit handles this by splitting his firms, not only by size of dividend surprise but also by the earnings change.

Other than these two aspects, the studies use methodology similar to that discussed earlier. Furthermore, their conclusions are similar. The market seems to adjust rapidly to new information.

The other finding of interest to investment professionals is that for a number of types of announcements, investigators have found a long-term drift in abnormal return (called postannouncement drift). For example, Agrawal, Jaffe, and Mandelker (1990) and Jaffe and Mandelker (1976) find that firms that acquire other firms have significant abnormal returns on average over the next five years. Similarly, Ritter (1983) studied initial public offerings and found that on average, new issues after the first day substantially underperform other securities on a risk-adjusted basis.

When we examine long-run abnormal returns, the choice of how expected return is measured is important, and there is significant controversy of whether the results of a long-term drift are real or the result of using the wrong model for measuring expected returns.

STRONG-FORM EFFICIENCY

In this section we discuss two issues. The first issue is whether insiders in their trading earn an excess return. Working at Atlantic Richfield, learning that your geologist had discovered massive oil fields off Alaska, and then trading on that information clearly leads to excess returns. It also likely leads to jail, as trading on inside information in the United States is illegal. Thus examining the profitability of insider trading is both an examination of the usefulness of insider information and the regulation of the Securities and Exchange Commission (SEC). The second issue is whether professional investors, security analysts, and mutual fund managers have profitable information.

Insider Trading

All investors who own more than a certain percentage of the outstanding shares or are at a sufficiently high management level are considered insiders. In the United States insiders

must list their purchases and sales with the SEC. If insiders trade on privileged information, then one would expect to see insiders purchase in months before the security price increases and sell in months before the security price declines. This pattern is, in fact, the pattern found by Jaffe (1974) and Lorie and Niederhoffer (1968). Furthermore, they found, using methodology similar to that discussed earlier, that insiders earned returns in excess of expected return. Unless these insiders just happened to possess superior analytical ability, their excess return must be due to the illegal exploitation of insider information.

Another indication of the usefulness of insider trading is a legal action involving the person who set the type for the Value Line forecasts. Value Line is an investment advisory service; it divides firms into five groups, depending on its estimate of next period's performance. The typesetter knew what the recommendations of Value Line would be before the paper was printed and sold these to two brokers at a large brokerage firm. The brokers, in turn, used it to manage money for their clients. As reported in the *Wall Street Journal* (1982), the brokers made a fair amount of money trading in the securities before they were apprehended.

Information in Analysts' Forecasts

Many authors have analyzed whether security analysts have information not incorporated into security prices. The majority of these studies suffer from selection bias and survivorship bias. Selection bias occurs because most studies analyze a set of historical analysts' forecasts, and access to these forecasts is controlled. Security analysts generally work for an investment organization that controls whether outsiders have access to prior analysts' forecasts. Furthermore, the investment organization is likely to systematically evaluate the forecasts of their analysts. The organizations that provide prior analysts' forecasts to academics are likely to be those where the organization knows normal evaluation techniques will show superior information. Thus, even if analysts had no information, academic studies would likely find information because the organizations supplying data for outside studies are the ones whose analysts by chance did well. We know of two studies that do not suffer from selection bias, since the forecasts they analyzed were prepared after the organizations to be studied were selected.[29]

These studies are by Dimson and Marsh (1984) and Elton, Gruber, and Grossman (1986).[30] Dimson and Marsh analyzed 4,000 return forecasts made for 200 of the largest U.K. common stocks provided by 35 different firms of analysts. The data were gathered by a large fund that requested their brokers to forecast excess return on shares assuming a zero excess return on the market (difference from the riskless rate). Dimson and Marsh correlated actual return with forecasted returns and found an average correlation coefficient of 0.08. This result is consistent with other research in the area. Recall that the square of the correlation coefficient is the percentage explained. Thus $(0.08)^2 = 0.0064$ of realized return is explained by analysts' forecasts of return. Forecasting ability differed across the 35 firms. The range of correlation coefficients for brokers recording more than 50 forecasts was -0.19 to $+0.26$. Furthermore, past forecasting ability was not predictive of future forecasting ability. The best estimate of which firm forecasted best in the next period was that all firms were equal. However, combining the forecasts did lead to improvement. The correlation between realized return and the average analyst's forecast was 0.12.

[29]The only possible bias was if the organizations with poor analysts refused to participate. In each case, however, the request was made by a major financial institution, so that no one refused to supply the information.

[30]Dimson and Marsh (1984) present an extensive bibliography and review of previous research on analysts' forecasts. The reader interested in further research in this area should consult their article.

The forecasts of return were utilized by a fund for actual trades. Despite the small amount of information contained in the forecasts, as indicated by the size of the correlation presented earlier, the performance of the fund exceeded the market by 2.2%. Tests showed that more than one-half of the information contained in the forecasts was incorporated into share price into the first month following the forecast. Thus a rapid reaction to analyst forecasts was necessary.

Elton, Gruber, and Grossman (1986) employed a database that was constructed by a large bank and disseminated under the name of I/B/O/S/S. This database contained the rankings of stocks into five groups: best buys, buys, holds, and two classes of sells. The data contained more than 10,000 classifications per month prepared by more than 720 analysts at 34 brokerage houses. An analysis of forecasts prepared in the form of discrete classifications is interesting because this is the form in which most decision makers in the financial community receive information. Elton, Gruber, and Grossman found that both a change in classification (e.g., from a hold to a buy or from a best buy to a buy) and the classification itself contained information. Excess risk-adjusted returns could be earned by buying upgraded stocks or stocks that were in a better classification, and selling downgraded stocks or stocks that were in a lower classification. Excess returns were found in the forecast (classification) month and for two months following the classification or change in classification.

Acting on changes in classification produced larger excess returns than acting on the recommendations themselves. In addition, no superior forecasters could be identified. One was better off following the advice of the average or consensus forecaster than the advice of any set of forecasters who performed best over a previous period.

Both Dimson and Marsh and Elton, Gruber, and Grossman find information in analysts' forecasts. There seems to be very little information about acting on the advice of single brokerage firms. By aggregating across brokerage firms, however, there appears to be real information that persists for short periods of time.

Publicly available analysts' information can suffer from selection bias and potentially suffers from survivorship bias. Survivorship bias occurs if the selection of the organization to be studied is based on knowledge concerning past forecasting skill. Survivorship bias can occur because one would expect that the firms that continue to be able to sell information to the public are those for which past information appears to be valuable. If analysts had no information, but by chance some were right in their forecasts and some were wrong, then a researcher who selected firms to study on the basis of currently existing firms and analyzes past data would likely find information in analysts' forecasts even if none existed.

Despite these problems, the most studied data on security analysts' information is the Value Line investment survey. As discussed earlier, Value Line publishes weekly rankings where securities are divided into five groups, with one being the firms with the best prospects and five the worst. Stickel (1985) analyzes the effect of a change in ranking using the event study methodology discussed earlier. He finds that prices change for those stocks that are moved from group 3 to group 2. For all stocks, the three-day price change averaged 2.44%, with the price change averaging 5.18% for small stocks. Furthermore, the price change was not reversed in subsequent periods. This is either additional evidence that analysts have information not fully incorporated into share price or confirmation that Value Line's reputation for having had good forecasts was confirmed.

Mutual Fund Performance

Dozens of researchers have examined the performance of mutual funds. A detailed discussion of mutual fund performance will be postponed to Chapter 25; however, a few comments will be made here.

Most of the studies evaluating mutual funds contain a serious survivorship bias in the sample analyzed. Putting together a sample of funds that exist today and then gathering historical data excludes funds that went out of business over the period studied. The funds that go out of business have below average performance. An investor purchasing a fund at the beginning of the period could potentially purchase a fund that disappears or survives. Because most studies look only at the performance of funds that survive, this makes performance look better than it actually is. Furthermore, because survivorship varies inversely with risk, analyzing a sample with survivorship bias will lead to high-risk groups appearing to have superior relative performance.

The performance of funds is clearly sensitive to the measure used to evaluate them. We know from prior sections that small stocks have excess returns when measured relative to the standard CAPM. Therefore, small stocks' managers would also show excess return relative to the standard CAPM even when small stocks' managers have no selection ability.

Studies that are survivorship free and measure performance relative to multiple indexes, such as those done by Elton, Gruber, Das, and Hvlarka (1990), find that managers underperform a combination of passive indexes combined to have the same risk as the fund being evaluated after management fees and expenses are taken into account. Furthermore, this underperformance is related to the management fees and expenses they charge. Thus mutual fund managers on average are unable to earn enough to compensate for the fees they charge and expenses they incur.

MARKET RATIONALITY

In the prior sections we discussed the speed with which information is incorporated into share price. We referred to this as *informational efficiency*. A number of authors are also concerned with whether prices accurately reflect investors' expectations about the present value of future cash flows. We will refer to this hypothesis as market rationality to distinguish it from informational efficiency, while recognizing that some authors use the word *efficiency* to apply to both ideas.

If markets exhibit rationality, there should be no systematic differences between share prices and the value of the security based on the present value of the cash flow to security holders. Much of the evidence on informational efficiency bears on market rationality. For example, if prices can be shown to respond to noneconomic variables such as stock splits, this would be powerful evidence against market rationality.

The existence of excess return as a function of firm characteristics and time patterns in security returns provides evidence against market rationality. Examples of these relationships include the size effect, the market/book effect, the January effect, and the day-of-the-week effect. For informational inefficiency it is necessary to show that a profitable trading strategy (including trading costs) can be constructed to exploit the anomaly. However, the mere presence of a persistent anomaly calls into question market rationality.

The major direct evidence on stock market rationality involves volatility tests, stock market crashes, and tests of market overreaction. Each will be discussed in turn.

Volatility Tests

Volatility tests examine the volatility of share prices relative to the volatility of the fundamental variables that affect share prices. Markets would be seen as irrational if share prices deviated a great deal more than variance in the fundamental variables affecting share prices would imply.

The volatility tests of LeRoy and Porter (1981) and Shiller (1981, 1984) are based on three assumptions:

1. Stock prices reflect the expectations of future dividends.
2. The real expected return on stock is constant over time.
3. Dividends can be described by a stationary process with a constant growth rate.

With these assumptions, they devise tests based on the volatility of real prices relative to the volatility of theoretical prices (determined by the present value of future dividends). They find that actual prices vary considerably more than theoretical prices and reject market rationality. The results found by LeRoy and Porter and Shiller have been reexamined by a number of authors. Marsh and Merton (1986) change the assumption of how dividends are determined, assuming that it is a positive function of past prices, and get results in direct opposition to those of Shiller.

Winners—Losers

DeBondt and Thaler (1985, 1987) have written several papers in which they argue that investors overreact. In particular, they find that stocks that are the most extreme losers have abnormally good subsequent performance and that stocks that have been the biggest winners have subsequent poor performance. They attribute this to overreaction on the part of investors. In particular, they construct portfolios each December of the 50 stocks that did the best and worst in the prior three or five years. They then measure performance in the subsequent three or five years. The portfolio of the 50 most extreme losers has high abnormal returns (especially in January), whereas the portfolio of 50 winners has negative abnormal returns.

Several aspects of the study are worth noting. First, this study is closely related to the tax-selling studies discussed elsewhere in this chapter. Second, the selection rule is exactly opposite the selection rule used in relative strength where past winners are selected. Third, one would expect that losers would have more small firms (partly because they are losers) than the winners category; we have discussed elsewhere the extra return of small firms in January.

The DeBondt and Thaler articles are an important challenge to market rationality and as such have received a fair amount of attention. Other authors have supported or refuted the finding. One area of controversy involves how expected return and thus abnormal return is calculated. Depending on the method of calculating expected return, evidence in support of DeBondt and Thaler (see Chopra, Lakonishok, and Ritter, 1992) or refute (see Ball and Khothari, 1990) is found. The second issue is how much of this effect is really another effect, such as the small firm effect or the tax-selling effect.[31]

Market Crash of 1987 and 2008

The stock market declined 23% in one day in October 1987. This decline followed a substantial decline on the prior Friday. For markets to be rational, people's expectations had to undergo substantial changes on Friday and Monday. Numerous researchers have tried to find news items that could have led to a major revision in expectations. Although there were clearly news items around the crash, it is hard to argue that they caused such a large

[31]There is also a survivorship issue, because the losers must still exist for five subsequent years; thus losers who went bankrupt are excluded.

change in expectations. Rather, panic, failure of the trading mechanism, and formula trading are usually given as reasons for the crash. The crash is not a challenge for informational efficiency unless one can show that it was predictable. It is, however, a greater challenge to market rationality. Nevertheless, it is possible that formula trading and market structure combined to allow a crash to occur, and once it occurred, people reevaluated their fundamental values because of the crash.

The crash of 2008 was precipitated by massive defaults in the subprime mortgage market. This was followed by a fall in housing prices, increased default on conventional mortgages, and a rise in unemployment. In the five-month period beginning in September 2008 and ending in February 2009, stock market prices fell by almost 50%. However, they grew by 36% from March 2009 until the end of 2009 and by 15% in the year 2010.

Is this a sign of market inefficiency only to the extent that an investor could have anticipated the results? Was the pessimism of 2008 a sign of market irrationality? Was the subsequent recovery a sign of market rationality or a sign of market irrationality? Has there been a structural change in U.S. and world economies? Has there been a structural change in average returns or the variances or covariances of returns in stock markets? Only time will tell.

CONCLUSION

Although it is difficult to summarize a chapter that discusses such a diverse set of literature, we will try.

The size of the abnormal return around announcement days is sufficiently large that any measure of expected return will show similar results. Thus the results of these studies are relatively insensitive to the measure chosen. These studies show that information is rapidly incorporated into share price and support efficient markets.

The results of studies of longer-term reaction such as postannouncement drift and the relationship of firm characteristics and abnormal returns depend on the model of expected return chosen. It is no coincidence that the implication of these studies for market efficiency is controversial.

Finally, the results of studies that find calendar patterns in security returns are inconsistent with market efficiency. However, the consistent finding of an inability of market professionals to outperform indexes raises questions as to the usefulness of these patterns.

QUESTIONS AND PROBLEMS

1. Discuss a trading strategy to utilize information such as that analyzed by Davies and Canes (1978). How low would transaction costs have to be for the rule to be profitable? How would risk affect the usefulness of the rule?

2. Filter rules are one way to use past price movements to predict future movements. Discuss an alternative way to use past data. How would you test this alternative?

3. One rule for selecting stocks that has been suggested is to buy high-growth, low-P/E stocks. How could this rule be tested?

4. It has been suggested that the EMH could be used to determine whether you have monopoly access to a type of information. Explain how this might be done.

5. If the market is semistrong-form efficient, must it be weak-form efficient?

6. You have been hired as a consultant to a large brokerage firm. The firm thinks it has discovered an inefficiency in the market. At certain times large blocks of stocks that are held by individuals and institutions under restrictive agreements become available

for trading. The date on which this happens is a matter of public record. How would you test whether the market is efficient with respect to the potential increased supply in stock?

7. A number of different models can be used to estimate return. Derive the circumstances under which the use of the zero-beta model might lead to the market being considered inefficient when the standard CAPM indicated efficiency.

8. Is the betting market at roulette an efficient market?

9. You have just become convinced that whenever the president of a company retires, an excess return can be made by buying the stock. Design a study to test this hypothesis.

BIBLIOGRAPHY

1. Asquith, P., and Mullins, David W. "Equity Issues and Offering Dilution," *Journal of Financial Economics*, **15** (1986), pp. 61–89.
2. ——. "The Impact of Initiating Dividend Payments on Shareholders' Wealth," *Journal of Business*, **56,** No. 1 (Jan. 1983), pp. 77–96.
3. Agrawal, A., Jaffe, J., and Mandelker, G. "The Post-merger Performance of Acquiring Firms: A Reexamination of an Anomaly," *Journal of Finance*, **47** (Dec. 1990), pp. 1605–1621.
4. Aharony, Joseph, and Swary, Itzhak. "Quarterly Dividend and Earnings Announcements and Stockholders' Returns: An Empirical Analysis," *Journal of Finance*, **35,** No. 1 (March 1980), pp. 1–12.
5. Albin, Peter. "Information Exchange in Security Markets and the Assumption of 'Homogeneous Beliefs,'" *Journal of Finance*, **XXIX,** No. 4 (Sept. 1974), pp. 1217–1227.
6. Amihud, Yakov, and Mendleson, Haim. "Asset Pricing and the Bid–Ask Spread," *Journal of Financial Economics*, **17,** No. 2 (Dec. 1986), pp. 223–250.
7. ——. "Liquidity, Asset Prices, and Financial Policy," *Financial Analysts Journal*, **47** (Nov./Dec. 1991), pp. 56–66.
8. Ang, Andrew, and Bekaert, Geert. "Stock Return Predictability: Is It There?" *Review of Financial Studies*, **20,** No. 3 (2007), pp. 651–707.
9. Ang, Andrew, Goetzmann, William, and Schaefer, Stephen. "Review of the Efficient Market Theory and Evidence: Implications for Active Investment Management," *Foundations and Trends in Finance*, **5,** No. 3 (2010), pp. 157–254.
10. Ball, Ray, and Kothari, S. "Anomalies in Relationships between Securities' Yields and Yield-Surrogates," *Journal of Financial Economics*, **6** (1978), pp. 103–126.
11. ——. "Nonstationary Returns: Implications for Tests of Market Efficiency and Serial Correlation in Returns," *Journal of Financial Economics*, **25,** No. 1 (Nov. 1989), pp. 51–74.
12. Banz, Rolf W. "The Relationship between Return and Market Value of Common Stock," *Journal of Financial Economics*, **9** (1981), pp. 3–18.
13. Banz, Rolf W., and Breen, William J. "Sample-Dependent Results Using Accounting and Market Data: Some Evidence," *Journal of Finance*, **41,** No. 4 (Sept. 1986), pp. 779–793.
14. Barry, Christopher B., and Brown, Stephen J. "Anomalies in Security Returns and the Specification of the Market Model," *Journal of Finance*, **39,** No. 3 (July 1984), pp. 807–818.
15. ——. "Differential Information and the Small Firm Effect," *Journal of Financial Economics*, **13** (1984), pp. 283–294.
16. Bar-Yosef, Sasson, and Brown, Lawrence. "A Reexamination of Stock Splits Using Moving Betas," *Journal of Finance*, **XXXII,** No. 4 (Sept. 1977), pp. 1069–1080.
17. Basu, S. "Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis," *Journal of Finance*, **XXXII,** No. 2 (June 1977), pp. 663–682.
18. ——. "The Relationship between Earnings' Yield, Market Value and the Return for NYSE Common Stocks: Further Evidence," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 129–156.

19. Berges, Angel, McConnell, John J., and Schlarbaum, Gary C. "The Turn-of-the-Year in Canada," *Journal of Finance*, **39,** No. 1 (March 1984), pp. 185–192.
20. Bernard, Victor L."Evidence that Stock Prices Do Not Fully Reflect the Implications of Current Earnings for Future Earnings," *Journal of Accounting and Economics*, **13,** No. 4 (Dec. 1990), pp. 305–340.
21. Bernard, Victor L., and Thomas, Jacob K. "Post-Earnings-Announcement Drift: Delayed Response or Risk Premium?" *Journal of Accounting Research*, **27** (Supplement, 1989), pp. 1–36.
22. Bhandari, Laxmi Chand. "Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence," *Journal of Finance*, **43,** No. 2 (June 1988), pp. 507–528.
23. Black, Fischer. "Yes, Virginia, There Is Hope: Tests of the Value Line Ranking System," *Financial Analysis Journal*, **29** (Sept./Oct. 1973), pp. 10–14.
24. Blanchard, Olivier, Rhee, Changyong, and Summers, Lawrence. "The Stock Market, Profit, and Investment," *Quarterly Journal of Economics*, **108,** No. 1 (Feb. 1993), pp. 115–136.
25. Blume, Marshall E., and Stambaugh, Robert F. "Biases in Computed Returns," *Journal of Financial Economics*, **12** (1983), pp. 387–404.
26. Bradley, Michael. "Interfirm Tender Offers and the Market for Corporate Control," *Journal of Business*, **53,** No. 4 (Oct. 1980), pp. 345–376.
27. Branch, Ben. "A Tax Loss Trading Rule," *Journal of Business*, **50,** No. 2 (April 1977), pp. 198–207.
28. Brauer, Gregory A. "Using Jump-Diffusion Return Models to Measure Differential Information by Firm Size," *Journal of Financial and Quantitative Analysis*, **21,** No. 4 (Dec. 1986), pp. 447–458.
29. Brenner, Menachem. "A Note on [57]: Risk, Return and Equilibrium: Empirical Tests," *Journal of Political Economy*, **84,** No. 2 (April 1976), pp. 407–409.
30. ——. "The Effect of Model Misspecification on Tests of the Efficient Market Hypothesis," *Journal of Finance*, **XXXII,** No. 1 (March 1977), pp. 57–66.
31. Brown, Philip, Keim, Donald B., Kleidon, Allan W., and Marsh, Terry A. "Stock Return Seasonalities and the 'Tax-Loss Selling' Hypothesis: Analysis of the Arguments and Australian Evidence," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 105–127.
32. Brown, Philip, Kleidon, Allan W., and Marsh, Terry A. "New Evidence on the Nature of Size Related Anomalies in Stock Prices," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 33–56.
33. Brown, Stephen J., and Warner, Jerold B. "Using Daily Stock Returns: The Case of Event Studies," *Journal of Financial Economics*, **14,** No. 1 (March 1985), pp. 3–32.
34. Brown, Stewart L., and Nichols, William D. "Assimilating Earnings and Split Information: Is the Capital Market Becoming More Efficient?" *Journal of Financial Economics*, **9,** No. 3 (Sept. 1981), pp. 309–314.
35. Campbell, John Y. "Stock Returns and the Term Structure," *Journal of Financial Economics*, **18** (June, 1987), pp. 373–399.
36. Campbell, John Y., and Shiller, Robert. "Stock Prices, Earnings and Expected Dividends," *Journal of Finance*, **43,** No. 3 (July 1988), pp. 661–676.
37. ——. "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, **1,** No. 3 (1988–1989), pp. 195–228.
38. Campbell, John Y., Grossman, Sanford J., and Wang, Jiang. "Trading Volume and Serial Correlation in Stock Returns," *Quarterly Journal of Economics*, **108,** No. 4 (1993), pp. 905–939.
39. Casatis, Patrick, Miles, James, and Woolridge, Randall. "Restructuring through Spinoffs: The Stock Market Evidence," *Journal of Financial Economics*, **33,** No. 3 (June 1993), pp. 293–313.
40. Chan, K. C. "On the Contrarian Investment Strategy," *Journal of Business*, **61,** No. 2 (April 1988), pp. 147–163.
41. Chan, K. C., and Chen, Nai-Fu. "An Unconditional Asset-Pricing Test and the Role of Firm Size as an Instrument Variable for Risk," *Journal of Finance*, **43,** No. 2 (June 1988), pp. 309–325.
42. ——. "Structural and Return Characteristics of Small and Large Firms," *Journal of Finance*, **46,** No. 4 (Sept. 1991), pp. 1467–1484.
43. Chan, K. C., Chen, Nai-Fu, and Hsieh, David A. "An Exploratory Investigation of the Firm Size Effect," *Journal of Financial Economics*, **14** (1985), pp. 451–471.

44. Chan, K. C., Hamao, Yasushi, and Lakonishok, Josef. "Fundamentals and Stock Returns in Japan," *Journal of Finance*, **46,** No. 5 (Dec. 1991), pp. 1739–1764.
45. Charest, Guy. "Dividend Information, Stock Returns, and Market Efficiency II," *Journal of Financial Economics*, **6,** No. 2/3 (June/Sept. 1978), pp. 297–330.
46. Chari, V. V., Jagannathan, Ravi, and Ofer, Aharon R. "Seasonalities in Security Returns: The Case of Earnings Announcements," *Journal of Finance Economics*, **21** (1988), pp. 101–121.
47. Cheng, L. Pao. "Reply to [189]," *Journal of Finance*, **XXVIII,** No. 3 (June 1973), pp. 742–745.
48. Chopra, N., Lakonishok, J., and Ritter, J. "Measuring Abnormal Performance: Do Stocks Overreact?" *Journal of Financial Economics*, **31,** No. 2 (April 1992), pp. 235–268.
49. Chordia, Tarun, and Swaminathan, Bhaskaran. "Trading Volume and Cross-Autocorrelations in Stock Returns," *Journal of Finance*, **55,** No. 2 (2002), pp. 913–935.
50. Christie, Andrew A., and Hertzel, Michael. "Capital Asset Pricing 'Anomalies': Size and Other Correlations," unpublished manuscript, University of Rochester (1981).
51. Ciccone, Stephen. J. "Investor Optimism, False Hopes and the January Effect." *Journal of Behavioral Finance* **12,** No. 3 (2011), pp. 158–168.
52. Cochrane, John H. "Volatility Tests and Efficient Markets: A Review Essay," *Journal of Monetary Economics*, **27,** No. 3 (June 1991), pp. 463–485.
53. Connolly, Robert A. "An Examination of the Robustness of the Weekend Effect," *Journal of Financial and Quantitative Analysis*, **24,** No. 2 (June 1989), pp. 133–169.
54. Conrad, Jennifer, and Kaul, Gautam. "Time-variation in Expected Returns," *Journal of Business*, **61,** No. 4 (Oct. 1988), pp. 409–425.
55. Conway, Delores A., and Reinganum, Marc R. "Stable Factors in Security Returns: Identification Using Cross-Validation," *Journal of Business and Economic Statistics*, **6** (1988), pp. 1–15.
56. Cootner, Paul. *The Random Character of Stock Market Prices* (Cambridge, MA: MIT Press, 1974).
57. Copeland, Thomas E., and Mayers, David. "The Value Line Enigma (1965–1978): A Case Study of Performance of Evaluation Issues," *Journal of Financial Economics*, **X,** No. 3 (Nov. 1982), pp. 289–322.
58. Corhag, Albert, Hawawini, Gabriel, and Michel, Pierre. "Seasonality in the Risk-Return Relationship: Some International Evidence," *Journal of Finance*, **42,** No. 1 (March 1987), pp. 49–68.
59. Damodaran, Aswath. "Economic Events, Information Structure, and the Return-Generating Process," *Journal of Financial Quantitative Analysis*, **20,** No. 4 (Dec. 1985), pp. 423–433.
60. Dann, Larry, Mayers, David, and Raab, Robert. "Trading Rules, Large Blocks, and the Speed of Price Adjustment," *Journal of Financial Economics*, **4,** No. 1 (Jan. 1977), pp. 3–22.
61. Davies, Peter Lloyd, and Canes, Michael. "Stock Prices and the Publication of Second-Hand Information," *Journal of Business*, **51,** No. 1 (Jan. 1978), pp. 43–56.
62. DeBondt, Werner F. M., and Thaler, Richard H. "Does the Stock Market Overreact?" *Journal of Finance*, **40** (July 1985), pp. 793–805.
63. ——."Further Evidence on Investor Overreaction and Stock Market Seasonality," *Journal of Finance*, **42,** No. 3 (July 1987), pp. 557–581.
64. Deo, Rohit S., and Richardson, Matthew. "On the Asymptotic Power of the Variance Ratio Test," *Econometric Theory*, **19,** No. 2 (2003), pp. 231–239.
65. Desai, Hemang, and Jain, Prem C. "An Analysis of the Recommendations of the 'Superstar' Money Managers at Barron's Annual Roundtable," *Journal of Finance*, **50,** No. 4 (Sept. 1995), pp. 1257–1273.
66. Dimson, Elroy, and Marsh, Paul. "An Analysis of Brokers' and Analysts' Unpublished Forecasts of UK Stock Returns," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1257–1292.
67. Divecha, Arjun, and Morse, Dale. "Market Responses to Dividend Increases and Changes in Payout Ratios," *Journal of Financial and Quantitative Analysis*, **18,** No. 2 (June 1983), pp. 163–173.
68. Dodd, Peter. "Merger Proposals, Management Discretion and Stockholder Wealth," *Journal of Financial Economics*, **8,** No. 2 (June 1980), pp. 105–137.

69. Dodd, Peter, and Ruback, Richard. "Tender Offers and Stockholders' Returns," *Journal of Financial Economics*, **5,** No. 3 (Dec. 1977), pp. 351–375.
70. Dryden, Myles. "A Source of Bias in Filter Tests on Share Prices," *Journal of Business*, **42,** No. 3 (July 1969), pp. 321–325.
71. Dyl, Edward. "Capital Gains Taxation and Year-End Stock Market Behavior," *Journal of Finance*, **XXXII,** No. 1 (March 1977), pp. 165–175.
72. Dyl, Edward A., and Maberly, Edwin D. "Odd-Lot Transactions around the Turn of the Year and the January Effect," *Journal of Financial and Quantitative Analysis*, **27,** No. 4 (Dec. 1992), pp. 591–604.
73. Eades, Kenneth M., Hess, Patrick J., and Kim, E. Han. "On Interpreting Security Returns During the Ex-Dividend Period," *Journal of Financial Economics*, **13** (1984), pp. 3–34.
74. Elton, Edwin J., Gruber, Martin J., and Busse, Jeff. "Do Investors Care about Sentiment?" *Journal of Business*, **71,** No. 4 (Oct. 1998), pp. 477–500.
75. Elton, Edwin J., Gruber, Martin J., Das, Sanjiv, and Hklarka, Matt. "Efficiency with Costly Information: A Reinterpretation of Evidence from Managed Portfolios," unpublished manuscript, New York University (1990).
76. Elton, Edwin J., Gruber, Martin J., and Grossman, Seth. "Discreet Expectational Data and Portfolio Performance," *Journal of Finance*, **XLI,** No. 3 (July 1986), pp. 699–712.
77. Emery, John. "The Information Content of Daily Market Indicators," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 2 (March 1973), pp. 183–190.
78. ——. "Efficient Capital Markets and the Information Content of Accounting Numbers," *Journal of Financial and Quantitative Analysis*, **IX,** No. 2 (March 1974), pp. 139–149.
79. Epps, Thomas. "Security Price Changes and Transaction Volumes: Theory and Evidence," *American Economic Review*, **LXV,** No. 4 (Sept. 1975), pp. 586–597.
80. ——. "Security Price Changes and Transaction Volumes: Some Additional Evidence," *Journal of Financial and Quantitative Analysis*, **XII,** No. 1 (March 1977), pp. 141–146.
81. ——. "Security Price Changes and Transaction Volumes: Additional Evidence: Reply to [179]," *American Economic Review*, **68,** No. 4 (Sept. 1978), pp. 698–700.
82. Eskew, Robert, and Wright, William. "An Empirical Analysis of Differential Capital Market Reactions to Extraordinary Accounting Items," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 631–674.
83. Fama, Eugene. "The Behavior of Stock Market Prices," *Journal of Business*, **38** (Jan. 1965), pp. 34–105.
84. ——. "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, **XXV,** No. 2 (March 1970), pp. 383–417.
85. ——. "Efficient Capital Markets II," *Journal of Finance*, **26,** No. 5 (Dec. 1991), pp. 1575–1617.
86. Fama, Eugene, and Blume, Marshall. "Filter Rules and Stock Market Trading," *Journal of Business*, **39** (Jan. 1966), pp. 226–241.
87. Fama, E., Fisher, L., Jensen, M., and Roll, R. "The Adjustment of Stock Prices to New Information," *International Economic Review*, **10,** No. 1 (Feb. 1969), pp. 1–21.
88. Fama, Eugene, and French, Kenneth R. "Permanent and Temporary Components of Stock Prices," *Journal of Political Economy*, **96** (April 1988), pp. 246–273.
89. ——. "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, **22,** No. 1 (Oct. 1988), pp. 3–25.
90. ——. "Business Conditions and Expected Returns on Stocks and Bonds," *Journal of Financial Economics*, **25,** No. 1 (Nov. 1989), pp. 23–49.
91. ——. "The Cross Section of Expected Stock Returns," unpublished manuscript, Graduate School of Business, University of Chicago (1991).
92. Fama, Eugene, and MacBeth, James. "Risk, Return and Equilibrium: Empirical Tests," *Journal of Political Economy*, **81,** No. 3 (May/June 1973), pp. 607–636.
93. Ferson, Wayne E., "Stock Market Regularities: A Synthesis of the Evidence and Explanations," in Elroy Dimson (ed.), *Stock Market Anomalies* (Cambridge: Cambridge University Press, 1988).
94. ——. "Trading Patterns, Bid-Ask Spreads, and Estimated Security Returns: The Case of Common Stocks at Calendar Turning Points," *Journal of Financial Economics*, **25,** No. 1 (Jan. 1989), pp. 75–97.

95. Ferson, Wayne E., and Harvey, Campbell R. "The Variation of Economic Risk Premiums," *Journal of Political Economy*, **99,** No. 2 (April 1991), pp. 385–415.
96. Finnerty, Joseph. "Insiders' Activity and Inside Information: A Multivariate Analysis," *Journal of Financial and Quantitative Analysis*, **XI,** No. 2 (June 1976), pp. 205–215.
97. ——. "Insiders and Market Efficiency," *Journal of Finance*, **XXXI,** No. 4 (Sept. 1976), pp. 1141–1148.
98. ——. "The Chicago Board Options Exchange and Market Efficiency," *Journal of Financial and Quantitative Analysis*, **XIII,** No. 1 (March 1978), pp. 29–38.
99. Firth, Michael. "The Information Content of Large Investment Holdings," *Journal of Finance*, **XXX,** No. 5 (Dec. 1975), pp. 1265–1281.
100. Flannery, Mark J., and Protopapdakis, Aris A. "From T-Bills to Common Stocks Investigating the Generality of Intra-Week Return Seasonality," *Journal of Finance*, **43,** No. 2 (June 1988), pp. 431–450.
101. Flavin, Marjorie A. "Excess Volatility in the Empirical Evidence," *Journal of Political Economics*, **91,** No. 6 (Dec. 1983), pp. 929–956.
102. Flood, Robert P., and Hodrick, Robert J. "Asset Price Volatility, Bubbles, and Process Switching," *Journal of Finance*, **41** (Sept. 1986), pp. 831–842.
103. French, Kenneth R. "Stock Returns and the Weekend Effect," *Journal of Financial Economics*, **8** (1980), pp. 55–70.
104. French, Kenneth R., Schwert, G. William, and Stambaugh, Robert F. "Expected Stock Returns and Volatility," *Journal of Financial Economics*, **19** (1987), pp. 3–29.
105. Friend, Irwin, and Lang, Larry H. D. "The Size Effect on Stock Returns: Is It Simply a Risk Effect Not Adequately Reflected by the Usual Measures?" *Journal of Business Finance*, **12,** No. 1 (March 1988), pp. 13–30.
106. Furst, Richard. "Does Listing Increase the Market Price of Common Stocks?" *Journal of Business*, **43,** No. 2 (April 1970), pp. 174–180.
107. Galai, Dan. "Tests of Market Efficiency of the Chicago Board Options Exchange," *Journal of Business*, **50,** No. 2 (April 1977), pp. 421–442.
108. Goetzmann, William N. "Patterns in Three Centuries of Stock Market Prices," *Journal of Business*, **4,** No. 1 (1993), pp. 249–270.
109. Goetzmann, William N., Ibbotson, Roger G., and Peng, Liang. "A New Historical Database for the NYSE 1815 to 1925: Performance and Predictability." *Journal of Financial Markets*, **4,** No. 1 (2001), pp. 1–32.
110. Goetzmann, William N., and Zhu, Ning. "Rain or Shine: Where Is the Weather Effect?" *European Financial Management*, **11,** No. 5 (2005), pp. 559–578.
111. Gibbons, Michael R., and Hess, Patrick J. "Day of the Week Effects and Asset Returns," *Journal of Business*, **54** (1981), pp. 579–596.
112. Givoly, Dan, and Ovadia, Arie. "Year-End Tax Induced Sales and Stock Market Seasonality," *Journal of Finance*, **38,** No. 1 (March 1983), pp. 171–185.
113. Givoly, Dan, and Pulman, Dan. "Insider Trading and the Exploitation of Inside Information: Some Empirical Evidence," *Journal of Business*, **58,** No. 1 (Jan. 1985), pp. 69–87.
114. Granger, C. W. "Some Aspects of the Random Walk Model of Stock Market Prices," *International Economic Review*, **9,** No. 2 (June 1968), pp. 253–257.
115. ——. "A Survey of Empirical Studies on Capital Markets," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
116. Granger, C. W., and Morgenstern, O. *Predictability of Stock Market Prices* (Boston: Heath, 1970).
117. Grier, Paul, and Albin, Peter. "Non-Random Price Changes in Association with Trading in Large Blocks," *Journal of Business*, **46,** No. 3 (July 1973), pp. 425–433.
118. Griffiths, R. J. "Relative Strength—An Indicator for Investment in the Equity Market," thesis, Department of Statistics, Cranfield College, Cambridge (1970).
119. Grinblatt, Mark, and Titman, Sheridan. "The Persistence of Mutual Fund Performance," *The Journal of Finance*, **47,** No. 5 (Dec. 1992), pp. 1977–1984.
120. Grossman, S. J., and Hart, O. D. "Disclosure Law and Takeover Bids," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 323–334.

121. Grossman, Sanford. "On the Efficiency of Competitive Stock Markets Where Trades Have Diverse Information," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 573–585.
122. Grossman, Sanford J., and Shiller, Robert J. "The Determinants of the Variability of Stock Market Prices," *American Economic Review*, **71,** No. 2 (May 1981), pp. 222–227.
123. Gultekin, Mustafa N., and Gultekin, N. Bulent. "Stock Market Seasonality: International Evidence," *Journal of Financial Economics*, **12** (1983), pp. 469–481.
124. Harris, Lawrence. "A Transaction Data Study of Weekly and Intradaily Patterns in Stock Returns," *Journal of Financial Economics*, **14** (May 1986), pp. 99–117.
125. Harvey, Campbell. "The World Price of Covariance Risk," *Journal of Finance*, **46,** No. 1 (March 1991), pp. 111–157.
126. Hausman, W. H., West, R. R., and Largay, J. A. "Stock Splits, Price Changes, and Trading Profits: A Synthesis," *Journal of Business*, **44,** No. 1 (Jan. 1971), pp. 69–77.
127. Hawkins, Eugene H., Chamberlin, Stanley C., and Daniel, Wayne E. "Earnings Expectations and Security Prices," *Financial Analyst Journal*, **40,** No. 5 (Sept./Oct. 1984), pp. 24–39.
128. Hirshleifer, David, and Shumway, Tyler. "Good Day Sunshine: Stock Returns and the Weather," *Journal of Finance*, **58,** No. 3 (2003), pp. 1009–1032.
129. Hodrick, Robert J. "Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement," unpublished manuscript, Northwestern University and National Bureau of Economic Research (1990).
130. Holthausen, Robert W., and Leftwich, Richard W. "The Effect of Bond Rating Changes on Common Stock Prices," *Journal of Financial Economy*, **17** (1986), pp. 57–89.
131. Huberman, Gur, and Kandel, Shmuel. "Value Line Rank and Firm Size," *Journal of Business*, **60,** No. 4 (Oct. 1987), pp. 577–589.
132. ——. "Market Efficiency and Value Line's Record," *Journal of Business*, **63,** No. 2 (April 1990), pp. 187–216.
133. Ibbotson, Roger, and Jaffe, Jeffrey. "Hot Issue Markets," *Journal of Finance*, **XXX,** No. 2 (Sept. 1975), pp. 1027–1042.
134. Jaffe, Jeffrey. "Special Information and Insider Trading," *Journal of Business*, **47,** No. 3 (July 1974), pp. 410–428.
135. Jaffe, Jeffrey, and Mandelker, Gershon. "The Fisher Effect for Risky Assets: An Empirical Investigation," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 447–458.
136. Jaffe, Jeffrey, and Westerfield, Randolph. "The Week-end Effect in Common Stock Returns: The International Evidence," *Journal of Finance*, **40,** No. 2 (June 1985), pp. 433–454.
137. Jaim, Prem C. "The Effect of Voluntary Sell-off Announcements on Shareholder Wealth," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 209–224.
138. ——. "Responses of Hourly Stock Prices and Trading Volume to Economic News," *Journal of Business*, **61,** No. 2 (April 1988), pp. 219–231.
139. James, Christopher, and Edmister, Robert O. "The Relation between Common Stock Return, Trading Activity and Market Value," *Journal of Finance*, **38,** No. 4 (Sept. 1983), pp. 1075–1086.
140. Jegadeesh, Narasimhan, and Titman, Sheridan. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, **48,** No. 1 (March 1993), pp. 65–91.
141. Jennergren, Peter. "Filter Tests of Swedish Share Prices," in Edwin J. Elton and Martin J. Gruber (eds.), *International Capital Markets* (Amsterdam: North-Holland, 1975).
142. Jennergren, Peter, and Korsvold, Paul. "The Non-Random Character of Norwegian and Swedish Stock Market Prices," in Edwin J. Elton and Martin J. Gruber, *International Capital Markets* (Amsterdam: North-Holland, 1975).
143. Jennings, Robert, and Starks, Laura. "Earnings Announcements, Stock Price Adjustment, and the Existence of Option Markets," *Journal of Finance*, **41,** No. 1 (March 1986), pp. 107–125.
144. Jensen, Michael. "Accounting Changes and Stock Prices," *Financial Analysts Journal*, **29,** No. 1 (Jan./Feb. 1973), pp. 48–53.
145. ——. "Some Anomalous Evidence Regarding Market Efficiency," *Journal of Financial Economics*, **6,** No. 2/3 (June/Sept. 1978), pp. 95–101.

146. Jensen, Michael, and Bennington, George. "Random Walks and Technical Theories: Some Additional Evidence," *Journal of Finance*, **XXV,** No. 2 (May 1970), pp. 469–482.
147. Jones, C. D., Pearce, O. K., and Wilson, J. W. "Can Tax-Loss Selling Explain the January Effect? A Note," *Journal of Finance*, **42,** No. 2 (June 1987), pp. 453–461.
148. Jovanovic, F. and Le Gall, P. "Does God Practice a Random Walk? The 'Financial Hysics' of a Nineteenth-century Forerunner, Jules Regnault." *European Journal of the History of Economic Thought*, **8** (2001), pp. 332–362.
149. Kamstra, Mark J., Kramer, Lisa A., and Levi, Maurice D. "Winter Blues: A SAD Stock Market Cycle," *American Economic Review*, **93,** No. 1 (2003), pp. 324–343.
150. Kaplan, Steven. "The Effect of Management Buyouts on Operating Performance and Value," *Journal of Financial Economics*, **24,** No. 2 (Oct. 1989), pp. 217–254.
151. Kato, K., and Shallheim, J. "Seasonal and Size Anomalies in the Japanese Stock Market," *Journal of Financial and Quantitative Analysis*, **20,** No. 2 (June 1985), pp. 243–260.
152. Katz, Steven. "The Price Adjustment Process of Bonds to Rating Reclassifications: A Test of Bond Market Efficiency," *Journal of Finance*, **XXIX,** No. 2 (May 1974), pp. 551–559.
153. Keim, Donald B. "Size Related Anomalies and Stock Return Seasonality Further Empirical Evidence," *Journal of Financial Economics*, **12** (1983), pp. 13–32.
154. ——. "Trading Patterns, Bid–Ask Spreads, and Estimated Security Returns: The Case of Common Stocks at Calendar Turning Points," *Journal of Financial Economics*, **25,** No. 1 (Nov. 1989), pp. 75–97.
155. Keim, Donald B., and Stambaugh, Robert F. "A Further Investigation of the Weekend Effect in Stock Returns," *Journal of Finance*, **39,** No. 3 (July 1984), pp. 819–840.
156. ——. "Predicting Returns in the Stock and Bond Markets," *Journal of Financial Economics*, **17,** No. 2 (Dec. 1986), pp. 357–390.
157. Kleidon, Allan W. "Bubbles, Fads and Stock Price Volatility Tests: A Partial Evaluation: Discussion," *Journal of Finance*, **43,** No. 3 (July 1988), pp. 656–659.
158. Kraus, Alan, and Stoll, Hans. "Price Impacts of Block Trading on the New York Stock Exchange," *Journal of Finance*, **XXVII,** No. 3 (June 1972), pp. 569–588.
159. Lakonishok, Josef, and Shapiro, Alan C. "Systematic Risk, Total Risk and Size as Determinants of Stock Market Returns," *Journal of Business Finance*, **10,** No. 1 (March 1986), pp. 115–132.
160. Lakonishok, Josef, Shleifer, Andrei, and Vishny, W. Robert. "Contrarian Investment, Extrapolation, and Risk," unpublished paper, University of Illinois (1993).
161. Lakonishok, Josef, and Smidt, Seymour. "Volume and Turn-of-the-Year Behavior," *Journal of Financial Economics*, **13** (1984), pp. 435–455.
162. ——. "Are Seasonal Anomalies Real? A Ninety-Year Perspective," *Review of Financial Studies*, **1** (Winter 1988), pp. 435–455.
163. Larcker, David F., Gordon, Lawrence A., and Pinches, George E. "Testing for Market Efficiency: A Comparison of the Cumulative Average Residual Methodology and Intervention Analysis," *Journal of Financial and Quantitative Analysis*, **XV,** No. 2 (June 1980), pp. 267–288.
164. Lo, Andrew W., and MacKinlay, A. Craig. "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, **1,** No. 1 (1988), pp. 41–66.
165. Larcker, David F., and Lys, Thomas. "An Empirical Analysis of the Incentives to Engage in Costly Information Acquisition: The Case of Risk Arbitrage," *Journal of Financial Economics*, **18** (1987), pp. 11–26.
166. Laurence, Martin M. "Weak Form Efficiency in the Kuala Lumpur and Singapore Stock Markets," *Journal of Business Finance*, **10** (Oct. 1986), pp. 431–445.
167. Lehmann, Bruce N. "Fads, Martingales, and Market Efficiency," *Quarterly Journal of Economics*, **105,** No. 1 (Feb. 1990), pp. 1–27.
168. LeRoy, Stephen F., and Porter, Richard D. "The Present-Value Relation: Tests Based on Implied Variance Bounds," *Econometrica*, **49,** No. 3 (1981), pp. 555–574.
169. Levy, Robert. "Relative Strength as a Criterion for Investment Selection," *Journal of Finance*, **22** (Dec. 1967), pp. 595–610.

170. Liu, Pu, Smith, Stanley D., and Syed, Azmat A. "Security Price Reaction to the *Wall Street Journal*'s Securities' Recommendations," *Journal of Financial and Quantitative Analysis*, **25,** No. 3 (Sept. 1990), pp. 399–410.
171. Lo, Andrew W., and MacKinlay, A. Craig. "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *Review of Financial Studies*, **1,** No. 1 (Spring 1988), pp. 41–66.
172. ——. "When Are Contrarian Profits Due to Stock Market Overreaction?" *Review of Financial Studies*, **3,** No. 2 (1990), pp. 175–205.
173. Logue, Dennis. "On the Pricing of Unseasoned Equity Issues: 1965–1969," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 1 (Jan. 1973), pp. 91–103.
174. Lorie, James, and Neiderhoffer, Victor. "Predictive and Statistical Properties of Insider Trading," *Journal of Law and Economics*, **11** (April 1968), pp. 35–53.
175. Malkiel, Burton G. "Returns from Investing in Equity Mutual Funds 1971 to 1991," *Journal of Finance*, **50,** No. 2 (June 1995), pp. 549–572.
176. Mandelbrot, Benoit. "Some Aspects of the Random Walk Model of Stock Market Prices," *International Economic Review*, **9,** No. 2 (June 1968), pp. 258–259.
177. ——. "When Can Price Be Arbitraged Efficiently? A Limit to the Validity of the Random Walk and Martingale Models," *Review of Economics and Statistics*, **LIII,** No. 3 (Aug. 1971), pp. 225–236.
178. Mandelker, Gershon. "Risk and Return: The Case of Merging Firms," *Journal of Financial Economics*, **1,** No. 4 (Dec. 1974), pp. 303–336.
179. Mankin, N. Gregory, Romer, David, and Shapiro, Matthew D. "An Unbiased Reexamination of Stock Market Volatility," *Journal of Finance*, **40,** No. 3 (July 1985), pp. 677–687.
180. Marsh, Terry A., and Merton, Robert C. "Dividend Variability and Variance Bounds Tests for the Rationality of Stock Market Prices," *American Economic Review*, **76,** No. 3 (1986), pp. 483–498.
181. Maxwell, William F. "The January Effect in the Corporate Bond Market: A Systematic Examination," *Financial Management*, **27,** No. 2 (1998), pp. 18–30.
182. Mech, Timothy S. "Portfolio Return Autocorrelation," *Journal of Financial Economics*, **34,** No. 3 (Dec. 1993), pp. 307–344.
183. Merton, Robert C. "On the Current State of the Stock Market Rationality Hypothesis," in Rudiger Dornbusch, Stanley Fisher, and John Bossons (eds.), *Macro-economics and Finance: Essays in Honor of Franco Modigliani* (Cambridge, MA: MIT Press, 1987).
184. Mitchell, Mark L., and Lehn, Kenneth. "Do Bad Bidders Become Good Targets?" *Journal of Political Economy*, **98,** No. 2 (April 1990), pp. 372–398.
185. Niederhoffer, Victor. "The Analysis of World Events and Stock Prices," *Journal of Business*, **44,** No. 2 (April 1971), pp. 193–219.
186. Neiderhoffer, Victor, and Osborne, M. F. M. "Market Making and Reversal on the Stock Exchange," *Journal of the American Statistical Association*, **61** (1966), pp. 897–916.
187. Niarchos, N. A. "Statistical Analysis of Transactions on the Athens Stock Exchange," thesis Nottingham College (1971).
188. Officer, Robert R. "Seasonality in the Australian Capital Markets: Market Efficiency and Empirical Issues," *Journal of Financial Economics*, **2** (1975), pp. 29–52.
189. Ohlson, James A., and Penman, Stephen H. "Volatility Increases Subsequent to Stock Splits: An Empirical Aberration," *Journal of Financial Economics*, **14,** No. 2 (June 1985), pp. 251–266.
190. Owens, Richard N., and Hardl, Charles O. *Interest Rates and Stock Speculation* (New York: Macmillan, 1995).
191. Oppenheimer, Henry R., and Schlarbaum, Gary G. "Investing with Ben Graham: An Ex Ante Test of the Efficient Market Hypothesis," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 3 (Sept. 1981), pp. 341–360.
192. Patel, Jayendu, Zeckhauser, Richard, and Hendricks, Darryll. "The Rationality Struggle: Illustrations from Financial Markets," *American Economic Review*, **81,** No. 2 (May 1991), pp. 232–236.

193. Patell, James M., and Wolfson, Mark A. "The Intraday Speed of Adjustment of Stock Prices to Earnings and Dividend Announcements," *Journal of Financial Economics*, **13,** No. 2 (June 1983), pp. 223–252.
194. Penman, Stephen H. "Insider Trading and the Dissemination of Firm's Forecast Information," *Journal of Business*, **55,** No. 4 (Oct. 1982), pp. 479–503.
195. ——. "The Distribution of Earnings News over Time and Seasonalities in Aggregate Stock Returns," *Journal of Financial Economics*, **18** (1987), pp. 199–228.
196. Pettit, R. Richardson. "Dividend Announcements, Security Performance, and Capital Market Efficiency," *Journal of Finance*, **XXVII,** No. 5 (Dec. 1972), pp. 993–1007.
197. ——. "The Impact of Dividend and Earnings Announcements: A Reconciliation," *Journal of Business*, **49,** No. 1 (Jan. 1976)), pp. 86–89.
198. Pettit, R. Richardson, and Westerfield, Randolph. "Using the Capital Asset Pricing Model and the Market Model to Predict Security Returns," *Journal of Financial and Quantitative Analysis*, **IX,** No. 4 (Sept. 1974), pp. 579–605.
199. Pinches, George. "The Random Walk Hypothesis and Technical Analysis," *Financial Analysts Journal*, **26,** No. 2 (March/April 1970), pp. 104–110.
200. Pinches, George, and Simon, Gary. "An Analysis of Portfolio Accumulation Strategies Employing Low-Priced Common Stocks," *Journal of Financial and Quantitative Analysis*, **VII,** No. 3 (June 1972), pp. 1773–1796.
201. Poterba, James M., and Summers, Lawrence H. "The Persistence of Volatility and Stock Market Fluctuation," *American Economic Review*, **76,** No. 5 (Dec. 1986), pp. 1142–1151.
202. Poterba, James, and Summers, Lawrence. "Mean Reversion in Stock Prices: Evidence and Implications," *Journal of Financial Economics*, **22,** No. 1 (Oct. 1988), pp. 27–59.
203. Praetz, Peter. "The Distribution of Share Price Changes," *Journal of Business*, **45,** No. 1 (Jan. 1972), pp. 49–55.
204. Regnault, Jules. *Calcul des Chances et Philosophie de la Bourse* (Paris: Mallet-Bachelier, 1863).
205. Reinganum, Marc R. "The Arbitrage Pricing Theory: Some Empirical Results," *Journal of Finance*, **37** (1981), pp. 27–35.
206. ——. "Misspecification of Capital Asset Pricing: Empirical Anomalies Based on Earnings Yields and Market Values," *Journal of Financial Economics*, **9** (March 1981), pp. 19–46.
207. ——. "The Anomalous Stock Market Behavior of Small Firms in January: Empirical Tests for Tax-Loss Selling Effect," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 89–104.
208. ——. "A Direct Test of Roll's Conjecture on the Firm Size Effect," *Journal of Finance*, **37,** No. 1 (March 1982), pp. 27–36.
209. ——. "The Anomalous Stock Market Behavior of Small Firms in January," *Journal of Financial Economics*, **12** (June 1983), pp. 89–104.
210. Reinganum, Marc R., and Shapiro, Alan C. "Taxes and Stock Return Seasonality: Evidence from the London Stock Exchange," *Journal of Business*, **60,** No. 2 (April 1987), pp. 281–295.
211. Ritter, Jay R. "The Buying and Selling Behavior of Individual Investors at the Turn of the Year," *Journal of Finance*, **43,** No. 3 (July 1983), pp. 701–717.
212. Roll, Richard. *The Behavior of Interest Rates: An Application of the Efficient Market Model to U.S. Treasury Bills* (New York: Basic Books, 1970).
213. ——. "A Possible Explanation of the Small Firm Effect," *Journal of Finance*, **36** (1981), pp. 879–888.
214. ——. "The Turn of the Year Effect and the Return Premium of Small Firms," *Journal of Portfolio Management*, **8,** No. 2 (1982), pp. 110–121.
215. ——. "On Computing Mean Returns and the Small Firm Premium," *Journal of Financial Economics*, **12** (1983), pp. 371–386.
216. Ross, Stephen A. "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, **13** (1976), pp. 341–360.
217. Rozeff, Michael S., and Kinney, William R., Jr. "Capital Market Seasonality: The Case of Stock Returns," *Journal of Financial Economics*, **3** (1976), pp. 379–402.

218. Rozeff, Michael S., and Zaman, Mir A. "Market Efficiency and Insider Trading: New Evidence," *Journal of Business*, **61,** No. 1 (Jan. 1988), pp. 25–44.
219. Scholes, Myron, and Williams, Joseph. "Estimating Betas from Nonsynchronous Data, *Journal of Financial Economics*, **5,** No. 3 (Dec. 1977), pp. 309–329.
220. Schultz, Paul. "Transaction Costs and the Small Firm Effect: A Comment," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 81–88.
221. Schwartz, Robert, and Whitcomb, David. "Evidence on the Presence and Causes of Serial Correlation in Market Model Residuals," *Journal of Financial and Quantitative Analysis*, **XII,** No. 2 (June 1977), pp. 291–313.
222. Schwert, G. William. "The Adjustment of Stock Prices to Information about Inflation," *Journal of Finance*, **36,** No. 1 (March 1981), pp. 15–30.
223. ——. "Size and Stock Returns, and Other Empirical Regularities," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 3–12.
224. Seyhum, N. Nejat. "The January Effect and Aggregate Insider Trading," *Journal of Finance*, **43,** No. 1 (March 1988), pp. 129–141.
225. Shiller, Robert J. "The Volatility of Long-term Interest Rates and Expectations' Models of the Term Structure," *Journal of Political Economy*, **87,** No. 6 (Dec. 1979), pp. 1190–1219.
226. ——. "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?" *American Economic Review*, **71,** No. 3 (June 1981), pp. 421–436.
227. ——. "Theories of Aggregate Stock Price Movements," *Journal of Portfolio Management*, **10,** No. 2 (Winter 1984), pp. 28–37.
228. ——. "The Marsh-Merton Model of Managers' Smoothing of Dividends," *American Economic Review*, **76,** No. 3 (June 1986), pp. 499–503.
229. ——. "Comovements in Stock Prices and Comovements in Dividends," *Journal of Finance* (1989).
230. Shleifer, Andrei. "Do Demand Curves for Stocks Slope Down?" *Journal of Finance*, **41,** No. 3 (July 1980), pp. 579–590.
231. Smidt, Seymour. "A New Look at the Random Walk Hypothesis," *Journal of Financial and Quantitative Analysis*, **III,** No. 3 (Sept. 1968), pp. 235–261.
232. Smirlock, Michael, and Starks, Laoura. "Day-of-the-Week and Intraday Effects in Stock Returns," *Journal of Financial Economics*, **17** (1986), pp. 197–210.
233. Solnik, Bruno. "Note on the Validity of the Random Walk for European Stock Prices," *Journal of Finance*, **XXVIII,** No. 5 (Dec. 1973), pp. 1151–1159.
234. Stickel, Scott E. "The Effect of Value Line Investment Survey Rank Changes on Common Prices," *Journal of Financial Economics*, **14,** No. 1 (March 1985), pp. 121–144.
235. Stoll, Hans R., and Whaley, Robert E. "Transaction Costs and the Small Firm Effect," *Journal of Financial Economics*, **12,** No. 1 (1983), pp. 57–79.
236. Summers, Lawrence H. "Does the Stock Market Rationally Reflect Fundamental Values?" *Journal of Finance*, **41,** No. 3 (July 1986). pp. 591–601.
237. Taylor, Stephen J. "Tests of the Random Walk Hypothesis against a Price-Trend Hypothesis," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1982), pp. 37–62.
238. Tinic, Seha M., Barone-Adesi, Giovanni, and West, Richard R. "Seasonality in Canadian Stock Prices: A Test for the Tax-Loss-Selling Hypothesis," *Journal of Financial and Quantitative Analysis*, **22,** No. 1 (March 1987), pp. 51–63.
239. Tinic, Seha M., and West, Richard R. "Risk and Return: January vs. the Rest of the Year," *Journal of Financial Economics*, **13** (1984), pp. 561–574.
240. Vermaelen, Theo. "Common Stock Repurchases and Market Signalling: An Empirical Study," *Journal of Financial Economics*, **9,** No. 2 (June 1981), pp. 139–183.
241. *The Wall Street Journal*, June 1982.
242. Watts, Ross. "The Information Content of Dividends," *Journal of Business*, **45,** No. 2 (April 1973), pp. 191–211.
243. ——. "Comment on [125]...On the Informational Content of Dividends," *Journal of Business*, **49,** No. 1 (Jan. 1976), pp. 81–85.

244. ——. "Comments on [153]...The Impact of Dividend and Earnings Announcements: A Reconciliation," *Journal of Business*, **49,** No. 1 (Jan. 1976), pp. 97–106.
245. Welch, Ivo, and Goyal, Amit. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction," *Review of Financial Studies*, **21,** No. 4 (2008), pp. 1455–1508.
246. West, Kenneth D. "Bubbles, Fads, and Stock Price Volatility Tests: A Partial Evaluation," *Journal of Finance*, **43,** No. 3 (July 1988), pp. 639–655.
247. West, Richard, and Tinic, Seha. "Portfolio Returns and the Random Walk Theory: Comment on [25]," *Journal of Finance*, **XXVIII,** No. 3 (June 1973), pp. 733–741.
248. Westerfield, Randolph. "The Distribution of Common Stock Price Changes: An Application of Transaction, Time and Subordinated Stochastic Models," *Journal of Financial and Quantitative Analysis*, **10,** No. 4 (Dec. 1977), pp. 743–765.
249. Yuan, Kathy, Zheng, Lu, and Zhu, Qiaoqiao. "Are Investors Moonstruck: Lunar Phases and Stock Returns," *Journal of Empirical Finance*, **13,** No. 1 (2006), pp. 1–23.
250. Zhang, Cherry Y., and Jacobsen, Ben. "Are Monthly Seasonals Real? A Three Century Perspective," *Review of Finance* (forthcoming, 2013).
251 Zarowin, Paul. "Does the Stock Market Overreact to Corporate Earnings Information?" *Journal of Finance*, **44,** No. 5 (Dec. 1989), pp. 1385–1399.

18

The Valuation Process

The search for the "correct" way to value common stocks, or even one that works, has occupied a huge amount of effort over a long period of time. Attempts have ranged from simple mechanical techniques for picking winners to hypotheses about the broad influences affecting stock prices. At one extreme, the attempt to find a simple rule for selecting stocks that will have above-average performance can be likened to the search for a perpetual motion machine. Just as the laws of thermodynamics tell us we cannot build a perpetual motion machine, the theory of efficient markets tells us there is no simple mechanical way to pick winners in the stock market, or at least none that will recover its cost of operation. Yet people continue to spend a disproportionate amount of time on both of these endeavors.

At the other extreme, the determinants of common stock prices are quite easy to specify in general terms. The price of common stock is a function of the level of a company's earnings, dividends, risk, the cost of money, and future growth rate. While it is easy to specify these broad influences, the implementation of a system that uses these concepts to successfully value or select common stocks is a difficult task. This is the task that a valuation model purports to accomplish.

A valuation model is a mechanism that converts a set of forecasts of (or observations on) a series of company and economic variables into a forecast of market value for the company's stock. The input to a valuation model is in terms of economic variables, for example, future earnings, dividends, variability of earnings, and so forth. The output is in terms of expected market value or expected return from holding the stock or, at the very least, a buy, sell, hold recommendation. The valuation model can be considered a formalization of the relationship that is expected to exist between a set of corporate and economic factors and the market's valuation of these factors.

Every financial organization employs a valuation model. Often the valuation model is implicit in the way the organization makes decisions rather than an explicit model. For example, the organization that holds an index fund is implicitly accepting the simple form of the capital asset pricing model (CAPM), though it may not explicitly invoke the model every time it makes a decision.[1] The company that buys low-price-earnings-ratio stocks is

[1]An index fund is a portfolio designed to replicate the market portfolio.

implicitly stating that only the present price earnings ratios, and not predictions of future growth or risk, affect the return that can be earned on stocks. The advantages of employing an explicit valuation model are tremendous. An explicit model requires the definition of relevant inputs. Furthermore, it ensures that these inputs will be systematically collected and used in a consistent manner over time. Finally, the use of a valuation model allows for feedback and control in the functioning of a financial institution. By breaking the process of portfolio analysis into forecasting inputs, valuing securities, and forming portfolios, the ability of the organization to perform in each of these areas can be measured, and those areas where the organization has ability can be capitalized upon.

For example, it is possible that an organization has a superior ability to forecast corporate variables but that the informational content of the forecasts is lost either in the valuation process or when securities are formed into portfolios. Only by breaking the process into logical steps can an institution see what it does well and what it does poorly. Only then can it capitalize on any special abilities it does have and improve its performance.[2]

In this chapter we review some of the more widely used approaches to security valuation. We have made no attempt to be exhaustive in the models we have selected. Rather we have attempted to present some typical models with perhaps some bias toward those that we find more appealing. We start this chapter with a review of the general discounted cash flow approach to security valuation.

DISCOUNTED CASH FLOW MODELS

Discounted cash flow models are based on the concept that the value of a share of stock is equal to the present value of the cash flow that the stockholder expects to receive from it.[3] We will argue that this is equivalent to the present value of all future dividends. To facilitate this argument, let us assume that a stockholder intends to hold a share of stock for one period. In this one period the stockholder will receive a dividend and the value of the stock when he or she sells it. If the dividend occurs at the end of the period, then the value of this share of stock should be given by

$$P_t = \frac{D_{t+1}}{(1+k)} + \frac{P_{t+1}}{(1+k)} \tag{18.1}$$

where

P_t = the price of a share at time t
D_{t+1} = the dividend received at time $t + 1$
P_{t+1} = the price at time period $t + 1$
k = the appropriate discount rate

To value this share, the stockholder must estimate the price at which the stock will sell one period hence. Using the method employed previously,

$$P_{t+1} = \frac{D_{t+2}}{(1+k)} + \frac{P_{t+2}}{(1+k)} \tag{18.2}$$

[2]We will have more to say about evaluation and control in Chapters 25 and 26.

[3]There is a long history of discussion in the academic literature about what should be discounted. Some authors argued earnings, some dividends, and others earnings plus noncash expenses such as depreciation. It turns out that, properly defined, these approaches are equivalent. See Miller and Modigliani (1961).

Substituting Equation (18.2) into Equation (18.1),

$$P_t = \frac{D_{t+1}}{(1+k)} + \frac{D_{t+2}}{(1+k)^2} + \frac{P_{t+2}}{(1+k)^2} \tag{18.3}$$

If we, in turn, solved for P_{t+2} and substituted in Equation (18.3), then solved for P_{t+3} and so on, we would find that

$$P_t = \frac{D_{t+1}}{(1+k)} + \frac{D_{t+2}}{(1+k)^2} + \frac{D_{t+3}}{(1+k)^3} + \cdots + \frac{D_{t+n+1}}{(1+k)^{n+1}} + \cdots \tag{18.4}$$

or that the value of share of stock is equal to the present value of all future dividends. Stating the problem in terms of a stream of dividends plus a terminal price as in Equation (18.3) does not avoid the problem of forecasting how the future price will be set. It is not incorrect to state the problem in this way, but it may confuse the real issue that dividends have (at least in theory) to be forecast into the indefinite future.[4]

At this point a question invariably arises: What happened to earnings? The reader instinctively feels that earnings should be worth something, whether they are paid out as dividends or not, and wants to know why they do not appear in the valuation equation. In fact, they do appear in the equation, but in the correct form. Earnings can be used for one of two purposes: they can be paid out to stockholders in the form of dividends, or they can be reinvested in the firm. If they are reinvested in the firm, they should result in increased future earnings and increased future dividends. To the extent earnings at any time, say, time *t*, are paid out to stockholders, they are measured by the term D_t, and to the extent they are retained in the firm and used productively, they are reflected in future dividends and should result in future dividends being larger than D_t. To discount the future earnings stream of a share of stock would be double counting because we would count retained earnings both when they were earned and when they, or the earnings from their reinvestment, were later paid to stockholders.

It might be worth noting that Equation (18.4), like any of the discounted cash flow (DCF) models discussed in this chapter, can be employed in any of three ways. First, P_t can be treated as the unknown and a value of P_t computed based on estimates of future dividends and the appropriate discount rate. This should be an estimate of the value of the stock, and a market price very different from value should be an indication that price will move in the direction of value.

Second, the present market price can be used for P_t, estimates of future dividends substituted in the equation and the equation solved for *k*. The value arrived at for *k* should be the rate of return the stockholder will earn on the stock.[5] If the value of *k* arrived at is higher than is warranted by the risk of the stock, then price should adjust upward and rates of return greater than *k* earned.[6]

---

[4]In practice, because of the discounting process, dividends that are expected to be received in the very distant future have very little impact on price.

[5]It should be obvious that the appropriate level of *k* is related to the risk of a stock. One way to determine the appropriate level of *k* is to employ capital market theory and to determine *k* from the security market line and an estimate of the firm's risk.

[6]If a stock is incorrectly priced, the rate of return earned may be different from the computed value of *k*. For example, if the value of *k* arrived at is higher than that warranted by the risk of the stock, the price of the stock should adjust upward. If this adjustment takes place rapidly, the return earned by buying the stock may be much greater than that implied by the computed value of *k*.

Finally, this equation can be converted to a price earnings ratio by simply dividing each side by earnings. The left-hand side of the equation would then represent the normal price earnings ratio at which the stock should sell.

To use an infinite dividend stream model in its purest form, it would be necessary to forecast the growth rate in dividends each year from now to infinity, use this infinite series of growth rates to derive a dividend stream, and then discount it back to the present. It is impractical to use the model in its purest form. No individual or institution can differentiate between short-term growth forecasts in the distant future. All users of infinite horizon DCF models make some simplifying assumptions about the pattern that growth will follow over time. A number of different assumptions about growth-rate patterns have been made and embodied in valuation models. We review a few of the more widely used ones here. In particular, we examine three sets of growth assumptions:

1. constant growth over an infinite amount of time[7]
2. growth for a finite number of years at a constant rate, then growth at the same rate as a typical firm in the economy from that point on[8]
3. growth for a finite number of years at a constant rate, followed by a period during which growth declines to a steady state level over a second period of years;[9] growth is then assumed to continue at the steady state level into the indefinite future

We can, for obvious reasons, refer to these three models respectively as one-period, two-period, and three-period growth models. It should be equally as obvious that we could have a four-period, five-period, or *N*-period growth model.

As we move down this list of models, we are assuming more complex growth patterns for a company. We may be gaining the potential to more accurately forecast what a company will do, but we are asking the analyst to supply not only more data but also data increasingly difficult to forecast. As the type of data we ask to have forecasted becomes more difficult and the amount of information grows, forecasts are likely to contain less information and more random noise. As models become more complex, a point of diminishing returns is reached. Where this point is cannot be answered in the abstract; it is a function of the forecasting skills of the organization employing the model. Thus the question can be answered only by examining the forecast ability of the organization that is using, or proposes using, one or more valuation models. Let us now turn to an examination of some of the DCF models mentioned earlier.

Constant Growth Model

One of the best-known and certainly the simplest DCF model assumes that dividends will grow at the same rate (g) into the indefinite future. If we define P_0 as today's price and D_1 as next period's dividend, the value of a share of stock is

$$P_0 = \frac{D_1}{(1+k)} + \frac{D_1(1+g)}{(1+k)^2} + \frac{D_1(1+g)^2}{(1+k)^3} + \cdots + \frac{D_1(1+g)^{N-1}}{(1+k)^N} + \cdots$$

[7]See Williams (1938) or Gordon (1962) for discussion of models of this type.

[8]See Malkiel (1963) for the presentation of a model of this type.

[9]See Molodovsky, May, and Chottinger (1965) for the presentation of a model of this type.

Using the formula for the sum of a geometric progression,[10]

$$P_0 = \frac{D_1}{k - g} \tag{18.5a}$$

This model states that the price of a share of stock should be equal to next year's expected dividend divided by the difference between the appropriate discount rate for the stock and its expected long-term growth rate. Alternatively, this model can be stated in terms of the rate of return on a stock as

$$k = D_1/P_0 + g \tag{18.5b}$$

The constant growth model is often defended as the model that arises from the following assumptions: the firm will maintain a stable dividend policy (keep its retention rate constant) and earn a stable return on new equity investment over time. If we let b stand for the fraction of earnings retained within the firm, r stand for the rate of return the firm will earn on all new investments, and I_t stand for investment at t, we get a very simple expression for growth. The formula requires an estimate of the growth in dividends over time. We can derive an expression for the growth in dividends by first examining the growth in earnings. Growth in earnings arises from the return on new investments. We can write earnings at any moment as

$$E_t = E_{t-1} + rI_{t-1}$$

If the firm's retention rate is constant, then

$$E_t = E_{t-1} + rbE_{t-1} = E_{t-1}(1 + rb)$$

Growth in earnings is the percentage change in earnings, or

$$g = \frac{E_t - E_{t-1}}{E_{t-1}} = \frac{E_{t-1}(1 + rb) - E_{t-1}}{E_{t-1}} = rb$$

Because a constant proportion of earnings is assumed to be paid out each year, the growth in earnings equals the growth in dividends, or

$$g_E = g_D = rb$$

[10]The sum of a geometric progression is given by Sum = First term [1 − (Common ratio)^N]/(1 − Common ratio), where N is the number of terms over which we are summing. For this model we have

$$P_0 = \frac{\frac{D_1}{1+k}\left[1 - \left(\frac{1+g}{1+k}\right)^N\right]}{1 - \frac{1+g}{1+k}}$$

As N goes to infinity and

$$\left(\frac{1+g}{1+k}\right)^N$$

goes to zero, we obtain the formula in the text.

Using this expression for growth, we can rewrite Equations (18.5a) and (18.5b) as[11]

$$P_0 = \frac{D_1}{k - rb} \qquad (18.6a) \qquad\qquad k = \frac{D_1}{P_0} + rb \qquad (18.6b)$$

It is worthwhile examining the implications of this model for the growth in stock prices over time. The growth in stock price is

$$g_P = \frac{P_{t+1} - P_t}{P_t}$$

Recognizing that P_t can be defined by Equation (18.6a) and that $P_t + 1$ is also given by Equation (18.5a), except that D_1 must be replaced by $D_1(1 + br)$, we find

$$g_P = br$$

Thus, under the one-period model, dividends, earnings, and prices are all expected to grow at the same rate. It might be worthwhile to point out the key role expectations about the future profitability of investment opportunities play in this model. The rate of return on new investments can be expressed as a fraction (perhaps larger than 1) of the rate of return security holders require:

$$r = ck$$

Substituting this in Equation (18.6b), noting that $D_1 = (1 - b)E_1$, and rearranging yields

$$k = \frac{(1-b)E_1}{(1-cb)P_0}$$

Notice that if the firm has no extraordinary investment opportunities ($r = k$), then $c = 1$, and the rate of return that security holders require is simply the inverse of the stock's price earnings ratio. On the other hand, if the firm has investment opportunities that are expected to offer a return above that required by the firm's stockholders ($c > 1$), the earnings price ratio at which the firm sells will be below the rate of return required by investors.[12]

Let us spend a moment examining how the single-period model might be used to select stocks. One way is to predict next year's dividends, the firm's long-term growth rate, and the rate of return stockholders require for holding the stock. Equation (18.5a) could then be solved for the theoretical price of the stock that could be compared with its present price. Stocks that have theoretical prices above their actual prices are candidates for purchase; those with theoretical prices below their actual price are candidates for sale. The same procedure could be followed using the equation in footnote 11 with respect to price earnings ratios.

Another way to use the DCF approach is to find the rate of return implicit in the price at which the stock is now selling. This can be done by substituting the current price, estimated dividend, and estimated growth rate into Equation (18.5a) and solving for the

[11]Analysts frequently like to work in terms of price earnings multiples. Because $D_1 = (1 - b)E_1$, if we divide both sides of Equation (18.5a) by earnings, we have

$$\frac{P_0}{E_1} = \frac{1-b}{k-br}$$

[12]For a detailed analysis of the role that investment opportunities play in the valuation of securities, see Elton and Gruber (1976).

discount rate that equates the present price with the expected flow of future dividends. If this rate is higher than the rate of return considered appropriate for the stock, given its risk, it is a candidate for purchase.

We illustrate the use of the single-period model with a simple example. In the past, xyz's stock was selling for $65 a share. At that time xyz's earnings were $3.99 per share, and it paid a $2.00 dividend. At that time a major brokerage firm was estimating xyz's long-term growth rate at 12% and its dividend payout rate at 50%. If we assume 13% is an appropriate discount rate of xyz, we would compute a theoretical price of

$$P_0 = \frac{2.00}{0.13 - 0.12} = \$200$$

While xyz's stock would seem to be undervalued selling at $65 a share, notice the sensitivity of this valuation equation to both the estimate of the appropriate discount rate and the estimate of the long-term growth rate. For example, if xyz's growth rate was estimated to be 9% rather than 12%, its theoretical price would be one-fourth as large, or $50.

The single-period model has the advantage of being the simplest of all the models we will examine. Furthermore, multiperiod growth models assume that after a number of years the firm grows at a constant rate forever. The one-period model derived in this section is used to determine firm value at the beginning of this constant growth period. Thus this simple model is used as a part of all subsequent models.

It seems logical to assume that firms that have grown at a very high rate will not continue to do so into the infinite future. Similarly, firms with very poor growth might improve in the future. While a single growth rate can be found that will produce the same value as a more complex pattern, it is so hard to estimate this single number, and the resultant valuation is so sensitive to this number, that many investment firms have been reluctant to use the single-period growth model. As a result, they have turned to two- and three-period growth models.

The Two-Period Growth Model

The simplest extension of the one-period model is to assume that a period of extraordinary growth (good or bad) will continue for a certain number of years, after which growth will change to a level at which it is expected to continue indefinitely.

The assumption that growth is constant after some point in time follows from the following line of reasoning. After some point in time (5 years, 10 years, 15 years) the analyst has no ability to differentiate between firms on the basis of growth. Many current high-growth firms will no longer have high growth, and many firms that are currently viewed as stodgy will be the dynamic high-growth firms of the future. Thus after some years, it is sensible not to differentiate between firms but simply to assume they all grow at the same rate. At this point the constant growth model is used.

Let us assume that the length of the first period is N years, that the growth rate in the first period is g_1, and that P_N is the price at the end of period N. We can write the value of the firm as[13]

$$P_0 = \left[\frac{D_1}{1+k} + \frac{D_1(1+g_1)}{(1+k)^2} + \frac{D_1(1+g_1)^2}{(1+k)^3} + \cdots + \frac{D_1(1+g_1)^{N-1}}{(1+k)^N} \right] + \frac{P_N}{(1+k)^N}$$

[13]Many authors write the first term's dividend as $D_0(1 + g_1)$. In this case, the dividend is the current dividend rather than next period's dividend.

This can, of course, be simplified using the formula for the sum of a geometric progression. The result is

$$P_0 = D_1\left[\frac{1-\left(\frac{1+g_1}{1+k}\right)^N}{k-g_1}\right]+\frac{P_N}{(1+k)^N}$$

In the two-period model we are assuming that after N periods, the firm exhibits a constant infinite growth. Thus the model developed in the earlier section describes P_N. If g_2 is the growth in the second period and D_{N+1} is the dividend in the $N + 1$ period, we have

$$P_N = \frac{D_{N+1}}{k-g_2}$$

The dividend in the $N + 1$ period can be expressed in terms of the dividend in the first period:

$$D_{N+1} = D_1(1+g_1)^{N-1}(1+g_2)$$

With these substitutions we have

$$P_0 = D_1\left[\frac{1-\left(\frac{1+g_1}{1+k}\right)^N}{k-g_1}\right]+\left[\frac{D_1(1+g_1)^{N-1}(1+g_2)}{(k-g_2)}\right]\left[\frac{1}{(1+k)^N}\right]$$

This formula can easily be solved for the theoretical price of any stock. However, the two-period model is often used in a slightly different form.

In one form of this model, in year N, the stock is assumed to change its characteristics so that it resembles the average stock in the economy. After year N, the stock is expected to grow at the same rate, have the same dividend policy, and be subject to the same risk as the average stock in the economy. In this case, the P/E ratio at which it sells in year N must be the same as the average P/E ratio for the economy. Let us define this as M_g.[14] The price in year N can then be defined as the expected earnings in year N times the appropriate P/E ratio or

$$P_N = \frac{P_N}{E_N}(E_N) = M_g E_N$$

If earnings grow at the same rate as dividends, then earnings in year N are next period's earnings E times $(1 + g_1)^{N-1}$, and price can be expressed as

$$P_0 = D_1\left[\frac{1-\left(\frac{1+g_1}{1+k}\right)^N}{k-g_1}\right]+\left[M_g E(1+g_1)^{N-1}\right]\left[\frac{1}{(1+k)^N}\right]$$

[14]See Malkiel (1963) for a formal derivation of this model.

Some rearrangement of the first term yields an expression that is more convenient to calculate:

$$P_0 = \frac{D_1}{k - g_1}\left[\frac{(1+k)^N - (1+g_1)^N}{(1+k)^N}\right] + M_g E(1+g_1)^{N-1}\left[\frac{1}{(1+k)^N}\right]$$

Notice that while we started with the value of a share of stock being equal to the present value of all future dividends, we could state the valuation in terms of the present value of a stream of dividends and terminal *N*-year earnings plus a terminal *P*/*E* ratio. While this has no mathematical advantages over the sum of an infinite stream of dividends, it does have the advantage of being expressed in terms with which the security analyst feels more at home.

Like the constant growth model, this type of model can be used to arrive at a theoretical price that can then be compared with actual price, or alternatively the rate of return implicit in the present price can be solved for. To illustrate the first of these calculations, let us return to our xyz example. Let us assume that the analyst expects xyz's 12% growth rate to continue for 15 years, after which the analyst expects xyz to become an average company. Furthermore, assume that after 16 years the *P*/*E* ratio for the market is expected to be 9.5. Then the theoretical value of xyz's stock would be[15]

$$P = \frac{2.00}{0.13 - 0.12}\left[\frac{(1.13)^{15} - (1.12)^{15}}{(1.13)^{15}}\right] + \frac{(9.5)(3.99)(1.12)^{14}}{(1.13)^{15}} = \$54.59$$

With a constant growth model, earnings, prices, and dividends all grow at the same rate. With two-period and three-period models, this is no longer true. With the model just described, dividends and earnings had two distinct growth rates. In the first period, dividends and earnings grow at g_1, and in the second period they grow at g_2. Price grows at neither. If g_1 is greater than g_2, then price grows initially at a rate above g_2 but below g_1 and declines to g_2. The longer the time of growth, the closer the original growth in price is to g_1.

As with all valuation models, the discount rate is the expected return on the stock if the price of the stock over time conforms to the valuation model. Table 18.1 illustrates these ideas for the xyz example discussed earlier.

This table was constructed as follows. First the price each year was computed by recalculating the price formula presented previously but by successively shortening the number of years for which extraordinary growth was expected to continue. For example, the entry for price opposite year 10 was found as follows:[16]

$$P_{10} = \left[\frac{2.00(1.12)^{10}}{(1.13)^1} + \frac{2.00(1.12)^{11}}{(1.13)^2} + \frac{2.00(1.12)^{12}}{(1.13)^3} + \frac{2.00(1.12)^{13}}{(1.13)^4} + \frac{2.00(1.12)^{14}}{(1.13)^5}\right] + \frac{9.5(3.99)(1.12)^{14}}{(1.13)^5}$$

$$P_{10} = [26.99] + 185.25(0.5427) = 127.55$$

[15]In the example we assume that both the \$2.00 dividend and the \$3.99 earnings will arise one period after the time of the valuation.

[16]This equation can also be written using the form shown in the text as

$$P_0 = \frac{2.00(1.12)^{10}}{(0.13 - 0.12)}\left[\frac{(1.13)^5 - (1.12)^5}{(1.13)^5}\right] + \frac{9.5(3.99)(1.12)^{14}}{(1.13)^5}$$

Table 18.1 Price and Dividend Behavior under a Two-Period Growth Model

	Price		Dollar Return		Percentage Return		
	At Beginning of Period 1	At End after Dividend Is Paid 2	Dividend at End of Period 3	Capital Gain 4	Dividend Yield 5	Price Appreciation 6	Total Return 7
0	54.58	59.68	2.00	5.10	3.66	9.34	13.00
1	59.68	65.20	2.24	5.52	3.75	9.25	13.00
2	65.20	71.17	2.51	5.97	3.85	9.16	13.01
3	71.17	77.61	2.81	6.44	3.95	9.05	13.00
4	77.61	84.55	3.15	6.94	4.06	8.94	13.00
5	84.55	92.02	3.52	7.47	4.16	8.84	13.00
6	92.02	100.03	3.95	8.01	4.29	8.70	12.99
7	100.03	108.61	4.42	8.58	4.42	8.58	13.00
8	108.61	117.78	4.95	9.17	4.56	8.44	13.00
9	117.78	127.55	5.55	9.77	4.71	8.30	13.01
10	127.55	137.92	6.21	10.37	4.87	8.13	13.00
11	137.92	148.89	6.97	10.97	5.05	7.95	13.00
12	148.89	160.45	7.79	11.56	5.23	7.76	12.99
13	160.45	172.58	8.77	12.13	5.47	7.50	12.97
14	172.58	185.25	9.77	12.67	5.66	7.34	13.00
15	185.25						

The expected dividend in year 11 was found by assuming the present dividend (2.00) would continue to grow at the 12% rate, giving it a value of $2.00 \times (1.12)^{10}$. Dividends from year 11 through year 15 are expected to grow each year at the 12% rate, and each dividend must be discounted back to year 10. The term in the large brackets is the value as of year 10 of the dividend received from year 11 to year 15. The last term is simply the price as of year 15 discounted back to year 10.

By successively employing this formula, we can arrive at the prices shown in columns 1 and 2 of Table 18.1. Dividends (column 3) are computed by applying the growth rate of 12% to the initial dividend. Once prices and dividends are computed, it is a simple matter to compute the percentage return from dividends and capital gains shown in columns 5 and 6. Adding the dividend return to the return from price appreciation, we get the total return shown in column 7.

A few concepts are made explicit by this example. First, note that the investor will get the 13% discount rate implied by our assumption, even though the contribution of capital gains and dividends to this return changes drastically over time. As the period of high growth draws to an end, more of the contribution comes from dividends and less from capital gains. In fact, if we examine growth in price for a moment, we can see an interesting pattern. We know that once steady state occurs, the single-period growth model is appropriate, and earnings, dividends, and price will grow at the same rate. For our example this rate can be found from

$$P_0 = \frac{D_1}{k - g_2}$$

where the price at the beginning of year 15 is \$185.25, $k = 0.13$, and the dividend is $\$2.00 \times (1.12)^{14}(1 + g_2)$. Solving for g_2 produces a value of 7.34%. Reexamining column 6 with this

number in mind shows that prices start growing at a rate in between the short- and long-run growth rates. The growth in price declines each period until the period of extraordinary growth is over and the growth in price equals the long-term growth rate in earnings and dividends.

Although we have chosen to present this model to solve for a theoretical price, the model could just as easily be used in a second way. The analysts would estimate all of the variables that enter the model, except the discount rate. The price used in the formula would be the current price. The formula can then be used to estimate the expected return.

An abrupt change from one growth rate to another for most stocks is probably not descriptive of reality. The three-period growth models discussed in the next section deal directly with this issue. Before discussing this model, however, a variation of the two-period growth model will be presented. It is perfectly feasible to allow the analyst to make specific forecasts of dividends for each year prior to the time that a steady-state growth rate is reached. This is a more detailed version of the model under discussion, for we don't impose a uniform growth rate for the first period. For illustration, assume that the analyst is willing to make forecasts for five years but after that does not wish to differentiate among firms.

Define D_t as the dividend at period *t*. If the constant growth model is used to value the firm from period 5 onward, then the price of a firm at zero is

$$P_0 = \frac{D_1}{(1+k)} + \frac{D_2}{(1+k)^2} + \frac{D_3}{(1+k)^3} + \frac{D_4}{(1+k)^4} + \frac{D_5}{\frac{k-g_2}{(1+k)^4}}$$

where g_2 is the growth rate after five years.

The analyst would explicitly forecast the first five dividends and then utilize long-term averages for the market to estimate g_2. Because most analysts view earnings as the fundamental valuable being forecast, dividends would likely be forecast by forecast earnings and payout ratios. Explicitly forecasting dividends until a period of steady growth allows a gradual change in growth rate. An alternative way to allow a gradual change in growth is to use a three-period model.

The Three-Period Model

The usual two-period model assumes that during the initial period, earnings would continue to grow at some constant rate. At year *N* the second period started and growth was assumed to drop instantly to some steady state value. Normally, the change to a new long-term growth rate would not occur instantly; rather, it would occur over a period of time. Thus a logical extension is to assume a third period. The resultant model would assume that in period 1, growth is expected to be constant at some level. The analyst must forecast both the level of growth and the duration of period 1. During period 2 the growth changes from its value in period 1 to a long-run steady state level. The analyst must forecast both the duration of period 2 and the pattern of change in growth. Although some firms (e.g., Wells Fargo) allow the analyst to select from among a predetermined set of patterns, most firms employ one pattern (usually linear) for all firms. The third and final period is the period of steady state growth. Many organizations assume that once a firm reaches steady state growth, it will have the same characteristics as the average firm in the economy. When this happens, the contribution of the third period to value can be found in a manner directly analogous to the formulation of the second period in the two-period model. Other users of the three-period model have assumed zero growth in the third period, whereas still others allow the analyst to forecast whatever growth is deemed appropriate.

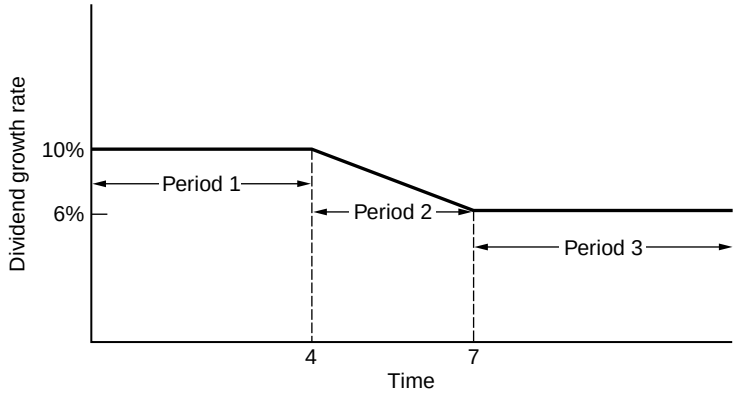


Figure 18.1 Growth rate pattern for a three-period model.

Figure 18.1 shows the growth rate in dividends for a typical three-period model. For the first four years, the firm is assumed to grow at a rate of 10%. After year 4, the growth rate of dividends is assumed to decline linearly to 6%. After year 7, the firm is assumed to grow at a rate of 6% forever. If we continue to assume a discount rate of 13%, the next dividend is \$2 and the next dividend payment is one year; hence the value of the firm would be

$$\text{Value} = \left[\frac{2}{(1.13)^1} + \frac{2(1.1)^1}{(1.13)^2} + \frac{2(1.1)^2}{(1.13)^3} + \frac{2(1.1)^3}{(1.13)^4}\right] + \left[\frac{2(1.1)^3(1.09)}{(1.13)^5} + \frac{2(1.1)^3(1.09)(1.08)}{(1.13)^6} + \frac{2(1.1)^3(1.09)(1.08)(1.07)}{(1.13)^7}\right] + \left[\frac{2(1.1)^3(1.09)(1.08)(1.07)(1.06)}{(0.13-0.06)}\right]\frac{1}{(1.13)^7}$$

The first two terms in brackets are the present value of the dividends received in the first and second growth period, respectively.

The last term in brackets is the value as of year 7 of this firm. This is simply an application of the constant growth model developed earlier. The numerator is the dividend as of year 8, one year from the valuation date. The denominator is the difference between the discount rate and the long-term constant growth rate of $(k - g)$. To calculate the impact of the value of the firm in year 7 on the firm today, simply discount the value back to the present by multiplying it by

$$\left[\frac{1}{1.13}\right]^7$$

In the preceding expression we chose to solve for the value of the stock. This value could then be compared with actual price to see if the stock should be purchased. An alternative way to employ this model is to set the right-hand side of the preceding equation equal to actual price but to leave the discount rate unspecified. The equation could then be solved for the discount rate implied by the analyst's expectation and the present price. This rate can be viewed as the analyst's estimate of expected return. The stock would be purchased or sold depending on the relationship between the analyst's estimate of expected return and what the firm considers a fair return for the stock given its risk.

As we move from a constant growth model to a two-period growth model to a three-period growth model, and perhaps even beyond this, we have increased the number and the complexity of the inputs the analyst must provide. If growth patterns are overly simplified, insufficient information will be provided by the forecasts. If they are made too complex, the forecasts are likely to be inaccurate. This trade-off is most apparent in the two extreme models discussed earlier. Analysts cannot develop year-by-year growth estimates into the indefinite future. At the other extreme, asking the analysts to provide only a single average growth forecast means losing the chance for the analyst to provide information about the future pattern of the company's growth. The trade-off between complexity and manageability will have to be made on the basis of the forecasting skills of an organization. No matter how this is decided, one of the principal benefits of using a valuation model can be the preparation of a comparable and explicit set of forecasts over time. Only if forecasts are made explicit can an organization evaluate and improve its performance over time.

Before leaving DCF models, it is worth noting another type of DCF model that is sometimes used by security analysts.

Finite Horizon Models

We have just seen that a model based on discounting a finite stream of dividends and a terminal price can be consistent with discounting an infinite stream of dividends. In this case, the finite nature of the model arose from consideration of future growth. Let us now look at a finite horizon model that arises from the way many organizations work, rather than from discounting an infinite stream of dividends.

Many organizations make short-run earnings forecasts for stocks (one- and two-year forecasts) and intermediate (five-year) growth forecasts. Analysts frequently predict future prices on P/E ratios rather than patterns of growth into the indefinite future. These forecasts can be incorporated into a valuation model by discounting expected dividends for the five years and the terminal price (the product of the expected P/E ratio and expected earnings based on the forecasted growth rate). Keep in mind that the five-year horizon used in this approach is not a function of the economics of the firm, the period over which a steady growth is expected to continue; rather, it arises from the forecasting pattern of the organization analyzing the stock. Although the model is mathematically equivalent to that discussed in the previous section, the rationale for the model is entirely different. Five years may not be an appropriate time horizon for the firm under study.

The major factor that separates this model from those we have previously discussed is the selection of a terminal P/E ratio without a specification of the economic rationale or assumptions behind either that P/E ratio or the five-year horizon. If the terminal P/E ratio is determined by assumptions about the future growth of the company, then the model reduces to one of those already discussed. If the terminal P/E ratio is simply asserted by the analyst based on experience or sense of the market, the analyst has implicitly made an assumption about the future growth pattern for the company. Assumptions about future growth cannot be avoided. If they are not made explicitly, they will be made implicitly by the selection of a terminal P/E ratio.

It would seem preferable to make growth assumptions explicitly rather than implicitly. If the analyst is going to use this type of model, he or she should at least explore the future growth rate implicit in the use of a terminal P/E ratio.

In fact, perhaps the most interesting aspect of this type of model is that it makes explicit the market expectations of future P/E ratios necessary to justify the price of a stock. That is, it can be used to answer the following question: given my estimate of both growth rates and the appropriate discount rate, what P/E ratio five years in the future justifies the present

price? Returning to our xyz example, we find that a *P*/*E* ratio of 16.50 would be necessary five years from now to justify the price of the stock today.[17]

The analyst could proceed to use one of the other growth models to discover the growth rate implicit in the expected future *P*/*E* ratio of 16.5. For example, using the constant growth rate assumption, we find that the implicit growth rate is close to 13% from year 5 into the indefinite future.

CROSS-SECTIONAL REGRESSION ANALYSIS

Although DCF models are enjoying a rapidly increased popularity in the investment community, they have been adopted by only a small fraction of the practicing security analysts.[18] The majority of security analysts still value common stocks by applying some sort of earnings multiple (price earnings ratio) to either present earnings, normalized earnings, or forecasted earnings. Approaches to the establishment of the *P*/*E* ratio cover a vast range. Some firms use the historical *P*/*E* ratios for companies or the historical *P*/*E* ratio for a company relative to the market *P*/*E* ratio. Another approach, and one popular in many of the standard texts of security analysis, is to list and discuss large numbers of factors that should affect *P*/*E* ratios but leave the weighting and often the explicit definition of these factors up to the security analyst.[19] Still another approach is to take the broad determinants of common stock prices, earnings, growth, risk, time value of money, and dividend policy and to measure these and weight them together in some manner to form an estimate of the *P*/*E* ratio. This section reviews one way to do this. We discuss the use of cross-sectional regression analysis to define the weights the market places on a set of hypothesized determinants of common stock prices. Attempts to use this technique to measure the influence of potential determinants of common stock prices were very popular in the 1960s, and there is an indication that interest in them has recently revived.

The relationship that exists in the market at any point in time between price or price earnings ratios and a set of specified variables can be estimated using regression analysis. This is the same tool that was used to determine betas in Chapter 7. Figure 18.2 presents the relationship between *P*/*E* ratios and forecasted growth for a sample of stocks as of the end of 1971. Each point in the diagram represents the *P*/*E* ratio and forecasted growth rate for a company as of the end of 1971. The straight line is fitted via regression analysis, and its equation is given by[20]

$$\text{Price/Earnings} = 4 + 2.3\ (\text{growth rate in earnings})$$

The usual technique of relating price or price earnings ratios to more than one variable is directly analogous to this. Called *multiple regression analysis*, it finds that linear combination of a set of variables that best explains price earnings ratios.

One of the earliest attempts to use multiple regression to explain price earnings ratios, which received wide attention, was the Whitbeck–Kisor model (1963). We indicated earlier

---

[17]This comes from an assumption of a constant payout ratio. The solution is

$$\$65 = P_0 = \frac{2.00}{1.13} + \frac{2(1.12)}{(1.13)^2} + \frac{2(1.12)^2}{(1.13)^3} + \frac{2(1.12)^3}{(1.13)^4} + \frac{2(1.12)^4}{(1.13)^5} + \left(\frac{P}{E}\right)(3.99)\frac{(1.12)^4}{(1.13)^5}$$

[18]For one survey in this area, see Bing (1971).

[19]Graham, Dodd, and Cottle (1962), perhaps the best-known book on security analysis, takes this approach.

[20]This example comes from Cohen, Zinbarg, and Zeikel (1973, p. 244).

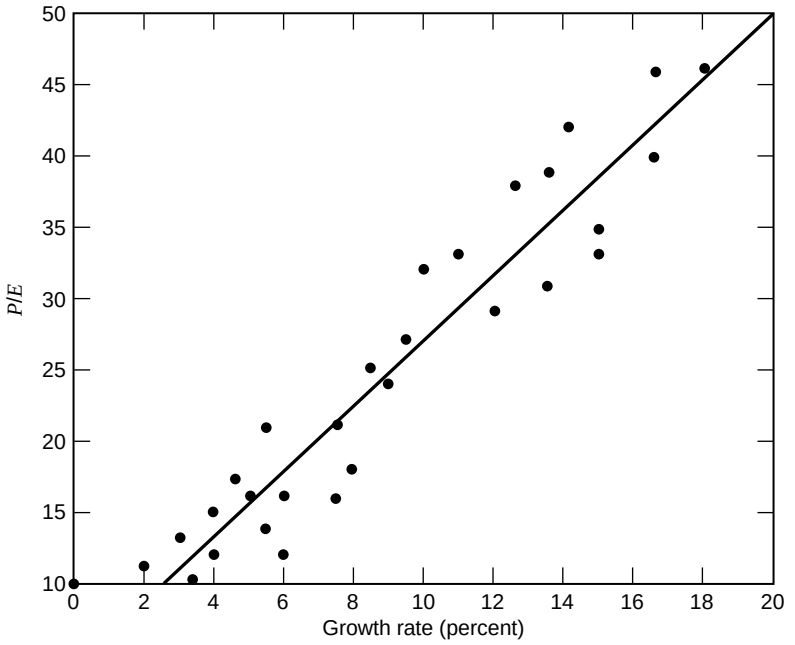


Figure 18.2 *P*/*E* ratios versus growth rates.

that the price of a share of stock was related to earnings, dividend policy, growth, and risk. We could have said, equally well, that the price earnings ratio of a stock was related to dividend policy, growth, and risk. It was exactly this relationship that Whitbeck and Kisor set out to measure. In particular, they obtained estimates of earnings growth rates, dividend payouts, and the variation (standard deviation) of growth rates from a group of security analysts. Then, using multiple regression analysis to define the average relationship between each of these variables and price earnings ratios, they found (as of June 8, 1962) that

$$\begin{aligned}\text{Price earnings ratio} = {} & 8.2 \\ & + 1.50\ (\text{earnings growth rate}) \\ & + 0.067\ (\text{dividend payout rate}) \\ & - 0.200\ (\text{standard deviation in growth rate})\end{aligned}$$

This equation represents the estimate at a *point in time* of the simultaneous impact of the three variables on the price earnings ratio. The numbers represent the weight that the market placed on each variable at that point in time. The signs represent the direction of the impact of each variable on the price earnings ratio. We might take some comfort from the fact that the signs are consistent with what theory and common sense would lead us to expect: the higher growth, the higher the dividends (growth held constant), and the lower risk, the higher the price earnings ratio. The equations tell us that, on average, a 1% increase in earnings growth is associated with a 1.5-unit increase in the price earnings ratio, a 1% increase in the dividend payout ratio is associated with a 0.067-unit increase in the price earnings ratio, and a 1% increase in the standard deviation of growth is associated with a 0.2-unit decrease in the *P*/*E* ratio.

An equation such as this can be used to arrive at the theoretical *P*/*E* ratio for any stock. Simply by substituting the forecasted earnings growth rate, dividend payout ratio, and risk for the stock on the right-hand side of the equation, one arrives at a theoretical *P*/*E* ratio. We can illustrate this with the xyz example used previously.[21] When xyz's price was $65, IBM's growth was forecast at 12%, its dividend payout ratio was 50%, and its standard deviation in growth rate was about 5. Substituting these numbers in the expression for price earnings ratios presented earlier, we get a theoretical *P*/*E* ratio of 28.55. Many researchers have taken what seems like a small step from here and advocated buying stocks with theoretical price earnings ratios above their actual price earnings ratios and selling short stocks with theoretical prices below their market price earnings ratio.

Literally hundreds of models like the Whitbeck–Kisor model have appeared in print since the 1960s. Every conceivable variable and combination of variables has been tried.[22] The common element of almost all of these models is that they are highly successful in explaining stock prices at a point in time, but they are much less successful in selecting the appropriate stocks to buy or sell short. It is not uncommon for these models to explain more than 80% of the difference in stock prices at a point in time. This gives us confidence that the models can be helpful in finding the variables and set of weights that determine price at a point in time. Why, then, haven't they been more successful in picking winners? The theory behind their use in finding under- and overvalued securities is that the market price will converge to the theoretical price before the theoretical price itself changes. There are at least three reasons why this might, in fact, not happen:[23]

1. Market tastes change. With changes in market tastes, the weight on each variable changes over time.
2. The values of the inputs, such as dividends and growth in earnings, change over time.
3. There are firm effects not captured by the model.

We discuss each of these in turn.

Market Tastes

One reason that price might not converge to theoretical price before theoretical price itself changes is that the parameters that determine theoretical price might change. Tastes, or the importance of certain variables in the market, change over time, and these changes are often rapid and drastic.

Let us return to the relationship between *P*/*E* ratios and growth examined earlier. The relationship found at the end of 1971 in a period of a bull market was

$$\text{Price earnings ratio} = 4 + 2.3 \text{ growth}$$

[21]The reader should be warned that this example is intended solely to illustrate the use of the model. The parameters of the model were estimated in 1962, and we are using 1976 data. The parameters of the model should not be expected to be stable over this period of time. We have more to say later about the stability of the parameters of regression models.

[22]Some of the more interesting models are Bower and Bower (1969), Gordon (1962), Gruber (1971), and Malkiel and Cragg (1970).

[23]A fourth reason should be briefly mentioned. The values in the equations are only estimates. Many researchers have tried large numbers of alternative definitions of variables or alternative variables in search of a "good" fit. Often what they are finding is spurious correlation (the variables happened to move together over the period). In this case there is no reason to believe the model will help select securities.

When the relationship was measured as of 1970 in a bear market using the same firms, it was

$$\text{Price earnings ratio} = 3 + 1.8 \text{ growth}$$

Notice that the importance of growth was higher in the bull market than it was in the bear market. For a stock with an expected growth rate of 20% per year, the estimated multiple rose from 39 in 1970 to 50 in 1971 or by more than 25%. The result is not surprising; it indicates the large magnitude of shift in market tastes that occurs over time. A similar shift with respect to a fuller set of determinants of common stock prices was reported by Gruber (1971). He examined the weight the market placed on dividends, growth, and three risk variables (earnings instability, financial leverage, and size) in each of 13 consecutive years. He found that the weights shifted drastically from year to year and were different at a statistically significant level.[24] Furthermore, the weights moved in a reasonable pattern, with growth becoming more important as the market moved up, and dividends less important. The opposite phenomenon occurred during downturns in the market. The shifts in the importance of the variables were more dramatic in those years when the market changed direction.

Input Data

Even if the market preference for variables remained stable over time, the theoretical value for a stock would change because the estimates of the variables like growth and dividends change. Input data are arrived at either by historical extrapolation or by the use of analysts' expectations. In any case, both the value of the inputs (earnings, growth, etc.) and expectations about these variables can and do change drastically over time. Every change in one of these variables—for example, expected growth—changes the theoretical value for a stock.

Firm Effects

Even when a model is constructed that explains a high fraction of the difference in stock prices, there are firms that have actual prices that lie above (or below) their theoretical prices and continue to do so period after period. Economists usually refer to this as *firm effects*. They are probably due to persistent influences that are not captured by the variables in the mode.[25] For example, in the early 1960s, tobacco stocks always had theoretical prices above their actual prices. This may well have been because theoretical prices did not take into consideration the threat of government intervention, while actual prices did.

Although cross-sectional regression models have been successfully used to examine the major determinants of common stock prices and the weight the market places on these determinants, the results from their use as a stock-selection tool have been mixed. Some authors, for example, Whitbeck and Kisor (1963), have reported an ability to outperform random selection; others, for example, Bower and Bower (1970) and Malkiel and Cragg (1970), have reported the failure of their models to lead to superior selection. The differences may be caused by the test periods used, the sample selected, or the authors' access to a better, or an inferior, set of forecasts.

[24]This means that it is inappropriate to pool cross-sectional samples in an attempt to define average weights.

[25]See Bower and Bower (1970) for a discussion of firm effects.

There is no doubt that cross-sectional regression models are helpful in understanding what has happened in the market over time. In addition, they may prove of some use in selecting stocks. However, the evidence at this time is not conclusive. It is clear that their usefulness is very dependent on the forecasting ability of the institution utilizing the model.

AN ONGOING SYSTEM

In this chapter we have considered several techniques for valuing common stock. The one with the strongest theoretical base involves the discounting of future dividends where the discount rate is appropriately formulated in terms of risk. In recent years several firms have attempted to implement stock valuation and selection systems that incorporate the DCF approach to stock selection and modern capital market theory. Perhaps the best known is the Wells Fargo stock evaluation system.

The first step in the Wells Fargo analysis system is to estimate the rate of return implicit in the price at which a stock is selling. They do this by finding the discount rate that equates the present value of all future dividends with price. The growth model they use to predict dividends is similar to the three-period model discussed earlier in the chapter.

In the Wells Fargo system, the analyst is required to estimate

1. dividends (and earnings per share) for each of the next five years
2. the fifth-year normalized earnings per share, growth rate, and payout
3. an eventual steady state payout and growth rate (the assumption here is that after a large number of years [larger than the five mentioned earlier], there will be a growth and payout rate that adequately describes the future behavior of the firm)
4. the number of years that are expected to elapse before the steady state condition is reached
5. the pattern of growth expected between the fifth year and the time that steady state growth is expected to begin; the analyst is free to select one from among several typical patterns that are presented to him

This gives an expected flow of dividends from the time of the analysis to infinity. This is used to find the expected rate of return, that is, the rate that equates expected dividends with present price.

In addition to dividend flows, the analyst provides estimates of the risk (beta) of each security. The analyst is given a measure of beta developed using historical data and is allowed to modify these estimates according to his analysis of the fundamental characteristics of the firm.

This results in an expected return and an expected beta for each stock. Now the expected return and expected beta for each of the companies in Wells Fargo's sample are plotted, and the straight line that fits these points is used as an estimated security market line.[26] Note that the Wells Fargo security market line is an expectational construct. Most security market lines (see Chapter 15) have been estimated on historical data (realizations) rather than expectations. The Wells Fargo security market line is a representation of a set of expectations. It represents the relationship between expected return and expected betas.

[26]Actually, Wells Fargo divides stocks into risk groups by beta range and examines the return on a stock against the average beta for its group.

If stock has a return (given its beta) above the security market line, it should offer a superior risk-adjusted return; if below, an inferior risk-adjusted return.

Perhaps an alternative explanation of this methodology might help. Analysts, by forecasting dividends, provide an estimate of the return expected from each stock. To understand whether this return is sufficient to compensate for risk, we need to know what the risk of the security is and the average expected rate of return the market requires for bearing that risk. The analyst estimates the risk (beta) on the stock. The average relationship between expected return and expected risk (the security market line) is found by looking at all stocks that analysts follow. If the stock offers a return above the return that should be warranted, given its risk (from the security market line), the stock should be a good buy. If it has a lower return, it should not be a good buy.

Let's consider a specific example. Table 18.2 shows a set of hypothetical data for a firm. It is divided into three periods. In the first period the analyst provides explicit forecasts of earnings and payout ratios for each year. The entries under the column dividends are then calculated by multiplying the first two columns. For the second period (the transitional period), the analyst forecasts the payout ratio (60%), the length of the period (three years), the long-term growth rate (6%), and the pattern of change in growth from the first period to the third period (linear). The growth rate in the first period averaged 10%. Given the analyst's forecasts, a growth rate of 9%, 8%, and 7% would be calculated for the three transitional years. The earnings in the last period are the earnings in period 8 compounded by the long-term growth rate of 6%. Finally, the stock price was assumed to be \$77.40. Thus the expected return is found by

$$\$77.40 = \frac{1.60}{(1+k)^1} + \frac{2.03}{(1+k)^2} + \frac{2.50}{(1+k)^3} + \frac{3.16}{(1+k)^4} + \frac{3.90}{(1+k)^5} + \frac{4.25}{(1+k)^6} + \frac{4.59}{(1+k)^7} + \frac{4.91}{(1+k)^8} + \left[\frac{5.21}{k-0.06}\right]^6 \frac{1}{(1+k)^8}$$

$$k = 0.10$$

Note that the first eight terms are the dividends shown in Table 18.2 discounted back to time zero. The last term has two parts, one in square brackets and the discount factor. The term in square brackets is the value of the firm as of period 8 using the constant growth model. Recall that the constant growth model is $D/k - g$. The D in the expression is the

Table 18.2 Forecasts for Company 1

Period	Year	Earnings	Payout Ratio	Dividends
First	1	4.00	40%	1.60
	2	4.50	45%	2.03
	3	5.00	50%	2.50
	4	5.75	55%	3.16
	5	6.50	60%	3.90
Transitional	6	6.50(1.09)	60%	4.25
	7	6.50(1.09)(1.08)	60%	4.59
	8	6.50(1.09)(1.08)(1.07)	60%	4.91
Final	9[a]	8.68	60%	5.21

[a]Earnings are 6.50(1.09) (1.08) (1.07) (1.06) = 8.68.

Table 18.3 Determining Mispriced Assets

Company	Expected Return	Beta	Excess Return
1	10%	1.2	−2.74
2	8	0.8	−1.86
3	15	1.4	+0.82
4	22	1.2	9.26
5	6	0.9	−4.58
6	18	1.6	2.38
7	16	1.8	−1.06
8	12	1.0	+0.70
9	4	1.2	−8.74
10	16	0.8	+6.14

dividend one period later than the time of valuation. For determining the value at period 8, the relevant dividend is the one paid at 9, whereas *g* is, of course, the 6% long-term growth rate. The value determined using the constant growth model is the value of the firm as of period 8. This value is discounted back eight periods to find the current value.

To decide if this company is a purchase or sale, the CAPM is used. The forecasted return of 10% determined by the dividend discount model is shown for company 1 in Table 18.3, together with the firm's beta. This beta is determined using the techniques discussed in Chapter 7, with the analyst allowed to modify it if she believes that the future beta is different from the best estimate using historical data.

Similar forecasts were made for companies 2 through 10 shown in Table 18.3. At this point, a CAPM is estimated by fitting a relationship to the data shown in Table 18.3. Running a least squares regression using the data shown in Table 18.3 results in

$$\overline{R}_i = 4.1 + 7.2\beta_i$$

Utilizing this equation to estimate equilibrium return allows the calculation of excess return for each stock that is shown in the last column. For example, for company 1, the expected return using the CAPM is

$$\overline{R}_i = 4.1 + 7.2(1.2) = 12.47$$

$$\text{Excess return} = 10 - 12.74 = -2.74$$

Thus the return of 10% that analysts forecast for company 1 is 2.74 below what is required given the risk of company 1, and company 1 would be a candidate for sale.

This approach has much to commend it. It uses the concept of the value of a share of stock being equal to the present value of future dividends, as well as the concepts of modern capital market theory. That is, it provides a consistent and theoretically defensible framework for the collection and use of output from security analysts. These are qualities that we have described earlier as being highly desirable.

Does this guarantee that the system will work? No. To be effective, the estimates from security analysts must contain real information. That is, their estimates of future dividends, future betas, and the security market line must, in combination, provide information about future returns.

Does the system work? An independent study done on the forecasting ability of the Wells Fargo Stock Advisory Service on the 250 stocks followed by TIAA CREF found that the service provided useful information on the relative value of stocks over a four-year

period. Although this is not conclusive, it does suggest that systematic use of the data supplied by security analysts can lead to superior performance.

An Evolving System of Security Selection

Just as the CAPM can be used as a tool in the stock selection process, the new models evolving from the arbitrage pricing theory (APT) and multifactor model literature can be used to enrich the stock selection process. Because the models employ more information about the process driving security returns, they allow for a more detailed structure for selecting stocks or designing stock selection systems. To illustrate some of the ways in which a multi-index model can be used, let us assume a particular return-generating process. We can think of this process as a simplified representation of the more detailed models described in Chapters 8, 14, and 16.

Assume

$$R_i = \bar{R}_i + b_{iD}I_D + b_{iy}I_y + b_{ip}I_p + b_{iO}I_O + e_i \tag{18.7}$$

where

R_i = the return on security i
$\bar{R}_i$ = the expected return on security i
I_D = innovations (unexpected changes) in default premiums
I_y = innovations in the difference between long- and short-term government securities
I_p = innovations (unexpected changes) in inflation
I_O = innovations in oil prices
b_{ij} = sensitivities to each of the influences generating returns
e_i = random error term

Under the assumptions of APT the equilibrium return for any asset should be the riskless rate of return (the T-bill rate) plus compensation for the sensitivity to different types of risks (b_{ij}s) inherent in the asset. Thus Equation (18.7) leads to

$$R_i = R_F + \lambda_D b_{iD} + \lambda_y b_{iy} + \lambda_p b_{ip} + \lambda_O b_{iO} \tag{18.8}$$

where the λs are the expected return for bearing sensitivity to each index (the market price of each type of risk). It is quite possible that a particular economic variable impacts returns over time but that the investor expects zero extra return for bearing sensitivity to it (the influence is not priced). For example, Roll and Ross (1980) found many more influences present in the return-generating process than were present in the equilibrium model (priced by the market).

There are a number of ways in which a model such as those depicted in Equations (18.7) and (18.8) can be used to manage a stock portfolio. The way in which the model can be used depends in part both on what an institution believes it can or cannot forecast and on the special characteristics of the institution's customer. We discuss each of these in turn.

Forecasting Ability

The simplest use of this model is analogous to the use of the CAPM to select securities, as discussed in the previous section of this chapter. In this section we assume that analysts can forecast the return on individual stocks but have no ability to forecast the risks (b_{ij}s)

of any security, future innovations in the macro variable driving the return-generating process (I's), or the market price of the factor influences (λs).[27]

The b_{ij}s for any stock can be estimated just like the β_is of the CAPM by using Equation (18.7) as a time-series multiple regression equation for each stock. Equation (18.8) can be run as a cross-sectional regression to estimate the λs by utilizing the analyst's forecast of the expected return for each stock and the estimate of the b_{ij}s for each stock arrived at from Equation (18.8). The results of this regression will be an equation such as

$$\overline{R}_i = 6 + 2b_{iD} + 0.8b_{iy} + 3b_{ip} \tag{18.9}$$

Note that in this example $\lambda_O = 0$, and thus the term $b_{iO}\lambda_O$ drops out of the equation. This implies that in equilibrium the investor does not believe the market gives an extra return to securities with sensitivity to oil prices. We made this assumption to reemphasize that some factors in the return-generating process need not be priced. This represents (just like the security market line) the *forecasted* equilibrium return on all stocks. For each security, an equilibrium return is determined by substituting the sensitivities (b_{ij}s) for that stock into Equation (18.9). The analyst then compares his forecasted return for each stock with the equilibrium return. If security analysts believe a stock will have a return above the equilibrium return determined by Equation (18.9), it is a good candidate to be purchased.

Although this use of the APT is directly analogous to what has become a popular use of the CAPM, there are other ways to employ this model in a forecast mode. The next most obvious use involves Equation (18.6a). We have assumed that the expected value of all innovations in the relevant macro variables is zero. Although this is true for the market as a whole, an institution may feel it can successfully forecast nonzero innovations in one or more of the macro variables. If so, it might choose to weight any portfolio it holds toward stocks with higher (or lower) sensitivities to these variables. For example, financial stocks have particularly high sensitivities with respect to the default variable. Therefore if a manager expected a large positive innovation in the default variable, he or she might want to overweight (or underweight, depending on the direction of the innovation) the fraction of financial stocks in the portfolio.

Similar forecasting arguments can be made with respect to differences between the manager's estimate of sensitivities (b_{ij}s) and consensus beliefs or differences in the manager's estimate of market price of any risks and the consensus beliefs.

All of these uses are based on an organization believing that it can forecast sensitivities (b_{ij}), market prices of risk (λ_j), or innovations (I_j) more accurately than the average or consensus belief that is incorporated in market prices. Even in the absence of differential forecasting belief, however, there is a potential use for the models under discussion.

Portfolios Customized for User Characteristics

Even if the institution employing the type of multi-index model under discussion does not believe that it possesses superior forecasting ability, it can take advantage of certain attributes of these models. The simplest use of this type of model is in the construction of index funds from a small number of stocks. Empirical evidence suggests that portfolios can be constructed that more closely mimic a target portfolio of securities (e.g., an index) when the target portfolio is matched with respect to several indexes rather than one.[28]

---

[27]Actually the system we describe involves implicit forecasts of the λs, but the forecasts result from a regression of the expected return forecast on each stock on the sensitivities rather than from a direct estimate.

[28]See Elton and Gruber (1989).

Replicating portfolios involving a small number of securities is especially useful in exploiting relative mispricing between markets. See, for example, the discussion of the use of futures on the Standard and Poor's (S&P) index in Chapter 24.

An even more appealing use of the APT is to construct portfolios that are suited to an individual customer's needs. For example, a pension fund that has future pension liabilities which are heavily sensitive to inflation might want to construct a portfolio that tends to give high payoffs when inflation is high. By examining the b_{ip} for individual stocks and industries, a portfolio can be constructed that resembles a market portfolio but that has returns that are more highly sensitive to inflation.

In deciding what position to take with respect to each source of risk portrayed in Equations (18.7) and (18.8), the investor is balancing extra risks against extra return. Although the market makes one particular trade-off, the investor may choose a different one because of her personal situation. As was just pointed out, an investor may choose to bear a high level of inflation risk because she has liabilities that are affected by inflation. Perhaps this can be seen most clearly by examining oil price risk. We assumed that the multiple regression resulted in an estimate of λ_O of zero (the investor believes that the oil price risk was not priced by the market). At first glance, we would conclude that no investor should hold a portfolio that has a value of b_{iO} other than zero. After all, there is no expected return from this risk, so why take it? But think of an investor whose total consumption expenditures are affected by oil prices. Such an investor will want to hold stocks with positive b_{iO} and bear this risk to cancel out some of the costs of consumption.

The use of an APT and multi-index model for stock selection is relatively new. We are only beginning to explore forecasting within the confines of the APT model. We are only starting to think about how the structure of the multi-index model can be used to design portfolios that have particular sets of multidimensional risk return characteristics that should appeal to specified groups of customers. This research, while still in its infancy, is promising.[29]

CONCLUSION

A valuation model can be considered as the black box that converts forecasts of fundamental data about companies and/or the economy into forecasts or evaluations of market price. In this chapter we have reviewed several approaches to valuation models. No valuation model can perform well if the forecasts on which it is based are of poor quality. On the other hand, good forecasts can be capitalized upon only if their effect on prices is evaluated in a sensible manner.

QUESTIONS AND PROBLEMS

1. A firm has just paid (the moment before valuation) a dividend of 55¢ and is expected to exhibit a growth rate of 10% into the indefinite future. If the appropriate discount rate is 14%, what is the value of the stock?
2. Consider the one-period growth model shown in Equation (18.5b). Assume the next period's dividend is $1, that stockholders require a 12% return, that new investment is expected to yield 14%, and that the retention rate is 50%. What is the implied fair price?

[29]See Elton and Gruber (1990, 1990) for an example.

3. Assume that price of the security discussed in Problem 2 was $30. Assume that all other information is the same except for the stockholders' required return. What does a $30 price imply for return?

4. Assume the information in Problem 2 and a price of $60. Furthermore, assume that the stockholder was most unsure concerning the return on new investment. How much would return have to change before the security was fairly priced?

5. The analyst who supplied you with the information in Problem 1 has just revised her forecast. She now realizes that the growth rate of 10% can continue for only five years, after which the company will have a long-term growth rate of 6%. Furthermore, at the end of the five years, she expects the company's payout rate to increase from its present 30% up to 50%. What value would you assign to the company?

6. Assume that the forecast for the company in Problem 5 was such that at the end of the fifth year its growth was to decline linearly for four years to reach the steady state 6% growth rate. Assume that the payout ratio was constant at 30% until it was changed to 50% at the end of the ninth year. What is the value of the company?

7. In Problem 2, assume that the price of the stock was $9 and solve for the expected rate of return from buying the stock.

8. In Problem 1, assume that the price of the stock was $9 and solve for the expected rate of return from buying the stock.

9. Consider the two-period model. Assume the same information as Problem 2, except that after 10 years, growth would change to 5%. What is the implied price?

10. Assume the security sold for $25, the two-period growth model is appropriate, and all other information is identical to Problem 9. What is the implied return?

11. Assume the same information as Problem 9. However, assume the length of time of the higher growth is uncertain. How long would it have to last to justify an $18 price?

12. Derive a three-period valuation model where the transitional period was N_2 years and involved a linear change from the first growth rate to a steady state growth rate.

BIBLIOGRAPHY

1. Altman, Ed. "Bankrupt Firm's Equity Securities as an Investment Alternative," *Financial Analysts Journal*, **25,** No. 4 (July/Aug. 1969), pp. 129–133.
2. Ambachtsheer, Keith. "Portfolio Theory and the Security Analyst," *Financial Analysts Journal*, **28,** No. 5 (Nov./Dec. 1972), pp. 53–57.
3. Arditti, Fred, and Pinkerton, John. "The Valuation and Cost of Capital of the Levered Firm with Growth Opportunities," *Journal of Finance*, **XXXIII,** No. 1 (March 1978), pp. 54–73.
4. Baker, Kent, and Haslem, John. "Toward the Development of Client-Specified Valuation Models," *Journal of Finance*, **XXIX,** No. 4 (Sept. 1974), pp. 1255–1263.
5. Baron, David. "Firm Valuation, Corporate Taxes, and Default Risk," *Journal of Finance*, **XXX,** No. 5 (Dec. 1975), pp. 1251–1264.
6. Baylis, Robert, and Bhirud, Suresh. "Growth Stock Analysis: A New Approach," *Financial Analysts Journal*, **29,** No. 4 (July/Aug. 1973), pp. 63–70.
7. Beaver, W., and Morse, D. "What Determines Price-Earnings Ratios?" *Financial Analysts Journal*, **34,** No. 4 (July/Aug. 1978), pp. 65–76.
8. Beidelman, Carl. "Pitfalls of the Price-Earnings Ratio," *Financial Analysts Journal*, **27,** No. 4 (Sept./Oct. 1971), pp. 86–91.
9. Bierman, Harold, and Hass, Jerome. "Normative Stock Price Models," *Journal of Financial and Quantitative Analysis*, **VI,** No. 4 (Sept. 1971), pp. 1135–1144.

10. Bierman, Harold, Downes, David, and Hass, Jerome. "Closed-Form Price Models," *Journal of Financial and Quantitative Analysis*, **VII,** No. 3 (June 1972), pp. 1797–1808.
11. Bildersee, John. "Some Aspects of the Performance of Non-convertible Preferred Stocks," *Journal of Finance*, **XXVIII,** No. 5 (Dec. 1973), pp. 1187–1201.
12. Bing, Ralph. "Survey of Practitioners' Stock Evaluation Methods," *Financial Analysts Journal*, **27,** No. 3 (May/June 1971), pp. 55–69.
13. Black, Fischer. "The Dividend Puzzle," *Journal of Portfolio Management*, **2,** No. 2 (Winter 1976), pp. 5–8.
14. Black, Fischer, and Scholes, M. "The Effects of Dividend Yield and Dividend Policy on Common Stock Prices and Returns," *Journal of Financial Economics*, **1,** No. 1 (May 1974), pp. 4–22.
15. Blume, Marshal Kraft, John, and Kraft, Arthur. "Determinants of Common Stock Prices: A Time Series Analysis," *Journal of Finance*, **XXXII,** No. 2 (May 1977), pp. 417–425..
16. Boness, James, Chen, Andrew, and Jatusipitak, Som. "On Relations among Stock Price Behavior and Changes in the Capital Structure of the Firm," *Journal of Financial and Quantitative Analysis*, **VII,** No. 4 (Sept. 1972), pp. 1967–1982.
17. Bower, Dorothy, and Bower, S. Richard. "Test of a Stock Valuation Model," *Journal of Finance*, **XXV,** No. 2 (May 1970), pp. 483–492.
18. Bower, Richard, and Bower, Dorothy. "Risk and the Valuation of Common Stock," *Journal of Political Economy*, **77,** No. 3 (May/June 1969), pp. 349–362.
19. Bower, Richard, and Wippern, Ronald. "Risk-Return Measurement in Portfolio Selection and Performance Appraisal Models: Progress Report," *Journal of Financial and Quantitative Analysis*, **IV,** No. 4 (Dec. 1969), pp. 417–447.
20. Breen, William, and Lerner, Eugene M. "Corporate Financial Strategies and Market Measures of Risk and Return," *Journal of Finance*, **XXVIII,** No. 2 (May 1973), pp. 339–351.
21. Breen, William, and Savage, James. "Portfolio Distributions and Tests of Security Selection Models," *Journal of Finance*, **XXIII,** No. 5 (Dec. 1968), pp. 805–819.
22. Brennan, M. J. "An Approach to the Valuation of Uncertain Income Streams," *Journal of Finance*, **XXVIII,** No. 3 (June 1973), pp. 661–674.
23. Brennan, Michael. "Valuation and the Cost of Capital for Regulated Industries: Comment on [30]," *Journal of Finance*, **XVII,** No. 5 (Dec. 1972), pp. 1147–1149.
24. Brigham, Eugene, and Pappas, James. "Rates of Return on Common Stock," *Journal of Business*, **42,** No. 3 (July 1969), pp. 302–316.
25. Chen, Nai-Fu. "Risk and Return of Value Stocks," *Journal of Business*, **71,** No. 4 (Oct. 1998), pp. 501–535.
26. Cohen, J., Zinbarg, E., and Zeikel, A. *Investment Analysis and Portfolio Management* (Homewood, IL: Richard D. Irwin, 1973).
27. Dennis, Charles N. "An Investigation into the Effects of Independent Investor Relations Firms on Common Stock Prices," *Journal of Finance*, **XXVIII,** No. 2 (May 1973), pp. 373–380.
28. Elton, Edwin J., and Gruber, Martin J. "The Effect of Share Repurchases on the Value of the Firm," *Journal of Finance*, **XXIII,** No. 1 (March 1968), pp. 135–149.
29. ——. "Marginal Stockholder Tax Rates and the Clientele Effect," *Review of Economics and Statistics*, **LII,** No. 1 (Feb. 1970), pp. 68–74.
30. ——. "Valuation and the Cost of Capital for Regulated Industries," *Journal of Finance*, **XXVI,** No. 3 (June 1971), pp. 661–670.
31. ——. "Valuation and the Cost of Capital for Regulated Industries: Reply to [21]," *Journal of Finance*, **XXVII,** No. 5 (Dec. 1972), pp. 1150–1155.
32. ——. "Asset Selection with Changing Capital Structure," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 3 (June 1973), pp. 459–474.
33. ——. "Valuation and the Asset Selection under Alternative Investment Opportunities," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 525–539.
34. ——. "Optimal Investment and Financing Patterns for a Firm Subject to Regulation with a Lag," *Journal of Finance*, **XXXII,** No. 5 (Dec. 1977), pp. 1485–1500.

35. ——. "A Multi-index Risk Model of the Japanese Stock Market," *Japan and the World Economy*, **1,** No. 1 (1988), pp. 21–44.
36. ——. "Expectational Data and Japanese Stock Prices," *Japan and the World Economy*, **1,** No. 4 (1990), pp. 391–401.
37. ——. "Portfolio Analysis with a Non-normal Multi-Index Return Generating Process," working paper, New York University (1990).
38. Elton, Edwin, Gruber, Martin, and Lieber, Zvi. "Valuation, Optimum Investment, and Financing for the Firm Subject to Regulation," *Journal of Finance*, **XXX,** No. 2 (March 1975), pp. 401–425.
39. Estep, Preston W. "A New Method for Valuing Common Stock," *Financial Analysts Journal*, **41,** No. 6 (Nov.–Dec. 1989), pp. 26–33.
40. Fewings, David. "The Impact of Corporate Growth on the Risk of Common Stocks," *Journal of Finance*, **XXX,** No. 2 (May 1975), pp. 525–531.
41. Foster, Earl. "Price-Earnings Ratio and Corporate Growth," *Financial Analysts Journal*, **26,** No. 1 (Jan.–Feb. 1970), pp. 96–99.
42. ——. "Price-Earnings and Corporate Growth: A Revision," *Financial Analysts Journal*, **26,** No. 3 (May–June 1970), pp. 115–118.
43. Fouse, W. "Risk and Liquidity: The Keys to Stock Price Behavior," *Financial Analysts Journal*, **32,** No. 3 (May–June 1976), pp. 35–45.
44. Fuller, Russel L., and Hsiu, Chi-Cheng. "A Simplified Common Stock Valuation Model," *Financial Analysts Journal*, **40,** No. 5 (Sept.–Oct. 1984), pp. 49–56.
45. Good, Walter. "Valuation of Quality-Growth Stocks," *Financial Analysts Journal*, **28,** No. 4 (Sept.–Oct. 1972), pp. 47–59.
46. Gordon, Myron. *The Investment, Financing, and Valuation of the Corporation* (Homewood, IL: Richard D. Irwin, 1962).
47. Graham, B., Dodd, D., and Cottle, S. *Security Analysis Principles and Techniques*, 4th ed. (New York: McGraw-Hill, 1962).
48. Granger, Clive W. J. "Some Consequences of the Valuation Model When Expectations Are Taken to Be Optimum Forecasts," *Journal of Finance*, **XXX,** No. 1 (March 1975), pp. 135–145.
49. Gruber, Martin J. *The Determinants of Common Stock Prices* (University Park: Pennsylvania State University Press, 1971).
50. Gupta, Manak. "Money Supply and Stock Prices: A Probabilistic Approach," *Journal of Financial and Quantitative Analysis*, **IX,** No. 1 (Jan. 1976), pp. 57–68.
51. Hakansson, Nils. "On the Dividend Capitalization Model under Uncertainty," *Journal of Financial and Quantitative Analysis*, **IV,** No. 1 (March 1969), pp. 65–87.
52. Hamburger, Michael, and Kochin, Levis. "Money and Stock Prices: The Channels of Influence," *Journal of Finance*, **XXVII,** No. 2 (May 1972), pp. 231–249.
53. Haugen, Robert. "Expected Growth, Required Return, and the Variability of Stock Prices," *Journal of Financial and Quantitative Analysis*, **V,** No. 3 (Sept. 1970), pp. 297–307.
54. Haugen, Robert, and Kumar, Prem. "The Traditional Approach to Valuing Levered-Growth Stocks: A Clarification," *Journal of Financial and Quantitative Analysis*, **IX,** No. 6 (Dec. 1974), pp. 1031–1044.
55. Haugen, Robert, and Pappas, J. L. "Equilibrium in the Pricing of Capital Assets, Risk-Bearing Debt Instruments, and the Question of Optimal Capital Structure," *Journal of Financial and Quantitative Analysis*, **VI,** No. 3 (June 1971), pp. 943–953.
56. ——. "Equilibrium in the Pricing of Capital Assets, Risk-Bearing Debt Instruments, and the Question of Optimal Capital Structure: A Reply," *Journal of Financial and Quantitative Analysis*, **VII,** No. 4 (Sept. 1972), pp. 2005–2008.
57. Haugen, Robert, and Udell, John. "Rates of Return to Stockholders of Acquired Companies," *Journal of Financial and Quantitative Analysis*, **VII,** No. 1 (Jan. 1972), pp. 1387–1398.
58. Hawkins, D. "Toward an Old Theory of Equity Valuation," *Financial Analysts Journal*, **33,** No. 6 (Nov.–Dec. 1977), pp. 48–53.

59. Hunt, Lacy. "Determinants of the Dividend Yield," *Journal of Portfolio Management*, **3,** No. 3 (Spring 1977), pp. 43–48.
60. Imai, Yutaka, and Rubinstein, Mark. "Equilibrium in the Pricing of Capital Assets, Risk-Bearing Debt Instruments, and the Question of Optimal Capital Structure: Comment," *Journal of Financial and Quantitative Analysis*, **VII,** No. 4 (Sept. 1972), pp. 2001–2003.
61. Jaffee, Jeffrey, and Mandelker, Gershon. "The Value of the Firm under Regulation," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 701–713.
62. Joy, Maurice, and Jones, Charles. "Another Look at the Value of *P*/*E* Ratios," *Financial Analysts Journal*, **26,** No. 4 (Sept.–Oct. 1970), pp. 61–64.
63. Keenan, Michael. "Models of Equity Valuation: The Great Bubble," *Journal of Finance*, **XXV,** No. 2 (May 1970), pp. 243–273.
64. Kummer, Donald, and Hoffmeister, Ronald. "Valuation Consequences of Cash Tender Offers," *Journal of Finance*, **XXXIII,** No. 2 (May 1978), pp. 505–515.
65. Latane, Henry, Joy, Maurice, and Jones, Charles. "Quarterly Data, Sort-Rank Routines, and Security Evaluation," *Journal of Business*, **43,** No. 3 (July 1970), pp. 427–438.
66. Litzenberger, Robert, and Budd, Alan. "Corporate Investment Criteria and the Valuation of Risk Assets," *Journal of Financial and Quantitative Analysis*, **V,** No. 4 (Dec. 1970), pp. 385–419.
67. Malkiel, Burton. "Equity Yields, Growth, and the Structure of Share Prices," *American Economic Review*, **53** (Dec. 1963), pp. 1004–1031.
68. ——. "The Valuation of Closed-End Investment-Company Shares," *Journal of Finance*, **XXXII,** No. 3 (June 1977), pp. 847–859.
69. Malkiel, Burton, and Cragg, John. "Expectations and the Structure of Share Prices," *American Economic Review*, **LX,** No. 4 (Sept. 1970), pp. 601–617.
70. Mehta, Dileep. "The Impact of Outstanding Convertible Bonds on Corporate Dividend Policy," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 489–506.
71. Miller, M., and Modigliani, F. "Dividend Policy, Growth, and the Valuation of Shares," *Journal of Business*, **34** (Oct. 1961), pp. 411–433.
72. Molodovsky, N., May, C., and Chottinger, S. "Common Stock Valuation," *Financial Analysts Journal*, **21** (March–Apr. 1965), pp. 104–123.
73. Myers, Stewart. "A Time-State Preference Model of Security Valuation," *Journal of Financial and Quantitative Analysis*, **III,** No. 1 (March 1968), pp. 1–33.
74. Nerlove, Marc. "Factors Affecting Differences among Rates of Return on Investments in Individual Common Stocks," *Review of Economics and Statistics*, **L,** No. 3 (Aug. 1968), pp. 312–331.
75. Roll, R., and Ross, S. A. "An Empirical Investigation of the Arbitrage Pricing Theory," *Journal of Finance*, **35,** No. 5 (Dec. 1980), pp. 1073–1103.
76. Sloane, William, and Reisman, Arnold. "Stock Evaluation Theory: Classification, Reconciliation, and General Model," *Journal of Financial and Quantitative Analysis*, **III,** No. 2 (June 1968), pp. 171–204.
77. Sorenson, Eric H., and Williamson, David A. "Some Evidence on the Value of Dividend Discount Models," *Financial Analysts Journal*, **41,** No. 6 (Nov.–Dec. 1985), pp. 60–69.
78. Warren, James. "A Note on the Algebraic Equivalence of the Holt and Malkiel Models of Share Valuation," *Journal of Finance*, **XXIX,** No. 3 (June 1974), pp. 1007–1010.
79. Whitbeck, V., and Kisor, M. "A New Tool in Investment Decision Making," *Financial Analysts Journal* (May–June 1963), pp. 55–62.
80. Williams, J.B. *The Theory of Investment Value* (Cambridge, MA: Harvard University Press, 1938).

19

Earnings Estimation

In the previous chapter we saw that both earnings and growth in earnings play a key role in valuation models. In this chapter we examine both the nature of earnings and some models for forecasting future earnings.[1]

We start this chapter by briefly reviewing some of the ambiguities associated with the term *earnings*. Different firms and even the same firm, at different times, can define earnings in alternative ways. A logical question is, If earnings can be defined differently, does the figure earnings per share, which shows up on the firm's income statement, have any impact on valuation? This question is examined in the second section of this chapter. As we will see, despite the ambiguous meaning of reported earnings, there is a real payoff from being able to forecast it.

The final two sections of this chapter examine models for forecasting future earnings. The first of the two sections examines the time series behavior of earnings, while the second discusses the relationship between earnings and other fundamental firm characteristics.

THE ELUSIVE NUMBER CALLED EARNINGS

The value of any asset is determined by its future earning power and not by what it cost at some time in the past. An economist would define earnings as cash flow plus the change in market value of an asset. Consider a bond originally purchased for $100 that carries a 10% interest rate. Assume the bond is worth $95 after one period. What has the earnings been on this investment over the period? The economist would say the earnings were $10 in interest plus the $5 decrease in value or a net of $5. An economist would apply the same principles to a physical investment. For example, if a manager purchases a machine, what are the earnings of the machine over the period? Clearly, one component of earnings is the profits earned from producing a product using the machine. An economist would argue that the change in the market value of the machine is also a part of earnings. The economist's concept of earnings is, of course, closely related to the idea of return we discussed in earlier chapters.

If an accountant reported the economist's definition of earnings and it was accurate, the analyst's job of valuing the asset or firm would be over. He or she could simply use the

[1]The authors wish to thank Paul Zarowin for help in revising this chapter.

estimate of the change in the value of the asset or firm together with the old selling price to determine the new price. However, there is circularity. The best estimate of the change in the value of the asset is, of course, the actual change in the value of the asset. If the actual change is determined by the accountant's estimate, then how can the actual change be used as the estimate? The accountant can still look at the fundamental characteristics of the firm and try to estimate the change in value. Similarly, the accountant could just try to report the income earned in the single period and leave the estimating of the change in value to others. The accounting profession pursues a policy somewhere in between. The number the accountant calls earnings is a mixture of the income earned and an attempt to measure some part of the change in the value of the asset. The accountant's treatments of depreciation, research and development expenditures, and pension liabilities have elements associated with them that are related to change in value. However, these attempts to measure changes in value tend to be related more to the allocation or using up of historic costs than they are to changes in market value. For example, depreciation reflects a somewhat arbitrary assumption about allocating the historical cost of an asset, as a change in value, over the life of an asset. The number the accountant uses for change in value (depreciation) is, at least in theory, related to the change in market value of an asset because the asset is used up. However, it makes no attempt to capture changes in the value of the asset due to either general or specific price changes. Thus accounting earnings are a mixture of the within-period earnings and an easily replicated but somewhat arbitrary allocation of some of the change in value of the asset. This is not the end of the story. There are still more difficulties with accounting earnings.

The most often cited problem is the lack of consistency in defining the components of earnings for different firms. Ashwinpaul Sondhi has prepared Table 19.1 to illustrate how, under current generally accepted accounting principles, a firm could show earnings of $1.98 or $4.41, depending on the choices made. Companies A and B are essentially the same company but have chosen different accounting methods in reporting income and cost.

The footnotes discuss in detail the differences between the assumptions made by the two companies. These include differences in assumptions concerning lives of investments, depreciation methods, pension costs, and forms of compensation.

Let us consider a few of these changes in detail. When the economy experiences very high inflation rates, the differences in the treatment of the costs of material used from inventory becomes important for many firms. There are two generally accepted methods of determining the cost of material used from inventory: LIFO and FIFO. LIFO (last in first out) uses as the cost of an item taken from inventory the cost of the last identical item purchased for inventory. FIFO (first in first out) uses as the cost of an item taken from inventory the cost of the oldest identical item in inventory. In periods of inflation the cost of the oldest item in inventory is often much lower than the cost of the most recent purchase of the same item. Using LIFO during periods of increasing inflation leads to lower reported earnings than the use of FIFO.

As a second case, consider pension liabilities. Pension liabilities are a source of increasing cost to firms. A firm makes a payment to the pension trustees to cover future liabilities. This payment is an expense to the firm and lowers earnings. The size of the payment a firm has to make depends, in part, on the assumed rate of return of the pension fund assets. Different rates of return can result in very different contributions and very different impacts on a firm's reported earnings.

There are no simple rules that allow an analyst to adjust the firm's earnings so that they are on a comparable basis. The impact of alternative accounting methods depends on the characteristics of the various firms. For example, the effect of differences in depreciation policies depends on the importance of fixed assets in the firm's costs, the age of the assets,

Table 19.1 Accounting Magic Using Generally Accepted Accounting Principles

	Company A	Adjustments ($ in 000s)	Company B
Sales revenue[a]	25,000	1,000	26,000
Other income:			
Equity/cost method affiliates[b]	1,500	(250)	1,250
Total revenues	26,500	750	27,250
Costs and expenses:			
Cost of goods sold	15,000		15,000
Selling, general and administrative	3,550		3,550
LIFO effect[c]	900	(900)	
Depreciation[d]	1,000	(400)	600
Amortization expense[e]	400	(250)	150
Exploration costs[f]	1,200	(550)	650
Pension costs[g]	750	(200)	550
Other postemployment costs[h]	300	(200)	100
Asset impairments[i]	300	(300)	
Compensation:			
Base salaries	400		400
Bonuses[j]	200	(200)	
Total costs/expenses	24,000	(3,000)	21,000
Pretax income	2,500		6,250
Tax expense[k]	525		1,837.5
Net income	1,975		4,412.5
Per share on 1,000 shares	1.975		4.4125

[a]*Revenue Recognition Methods*

Firms have considerable latitude concerning when they recognize revenue. For example, variations can occur because of differences in estimates of the degree and cost of completion in long-term construction contracts where the percentage-of-completion method is used, in revenue recognition of installment sales, and in the use of sales-type or direct-financing leases where operating leases should be used by the lessor. The impact of different revenue recognition is $1,000 more in income recognized by Company B. We assume this amount has not yet been received in cash; however, deferred taxes must be reported.

[b]*Equity/Cost Method Affiliates*

Company A owns 20% or more of the voting common stock in another company. Its proportionate share of the earnings of this investment, $1,500, is reported as a component of other income. It is assumed that Company A has received dividends of $1,250 from the investment. Company B owns less than the 20% threshold and cannot use the equity method. Under the cost method, it reports as other income the $1,250 received as dividends, but it does not report the additional $250 that would have been recorded under the equity method. This is the one instance where Company A and B are slightly different. It is included because it is an important difference between companies.

The investment in other companies also affects taxes. Eighty percent of dividends received from other corporations are tax-exempt. Thus Company A recognizes a tax expense and liability for 20% of the $1,250 received in dividends. Assuming a 35% corporate tax rate, the tax liability is $87.50 or [0.20 $\times$ $1250 $\times$ 0.35]. Company B recognizes the same tax expense and liability on the $1,250 received as dividends. However, deferred taxes must be recorded on the additional $250 recorded by Company A. Company A may assume that it will receive this amount as dividends or capital gains in the future. We use the latter because it is more conservative. Company A records an additional tax expense and deferred taxes payable. Because for corporations that tax is payable on the full amount of the capital gain, the tax is $87.50 or [0.35 $\times$ 250].

Note: Companies may assume either indefinite reinvestment of undistributed earnings ($250 in this case) or that these earnings will be received in a tax-free liquidation. Both assumptions would allow the company to record the $250 as income without any tax impact. This election is available only for companies with more than a 50% ownership share.

(continues on next page)

[c]*LIFO Effect*
Company A uses the LIFO inventory valuation method and Company B uses FIFO. In periods of increasing prices and stable or increasing inventories, LIFO firms will report higher cost of goods sold. The difference between FIFO and LIFO cost of goods sold is the LIFO effect, $900 in this case.
[d]*Depreciation*
Company A uses accelerated depreciation methods with shorter lives, whereas Company B uses straight-line depreciation with longer lives.
[e]*Amortization Expense*
Company B amortizes goodwill, patents, and copyrights over the maximum periods allowed, whereas Company A uses shorter lives.
[f]*Exploration Costs*
Company B capitalizes all exploration costs, whereas Company A expenses dry-hole costs.
[g]*Pension Costs*
The difference in pension costs, $200 lower for Company B, is due to a difference in assumptions of discount rates, assumed rates of return on assets, and different allocation to expense of the difference between actual results and actuarial assumptions.
[h]*Other Postemployment Contracts*
Costs of health care and life insurance benefits promised to employees are recognized as incurred by Company B. Current accounting does not require accrual of these costs (as in pensions). However, Company A has recorded current costs and accrued future costs. The values we used may understate the true differences, because studies have reported accrued amounts of as much as 20–30 times the periodic cost.
[i]*Asset Impairments*
Accounting guidelines for the timing and measurement of impairments (loss in value) to long-lived assets are at best vague and inconsistently applied. Firms also have considerable discretion with respect to reporting impairments to the carrying values of receivables, marketable securities, and investments in affiliates. Here Company A has recognized loss due to impairment, whereas Company B has not yet done so.
[j]*Compensation*
Company B uses stock options for bonuses, whereas Company A pays them in cash.
[k]*Tax*
For both companies that tax is the tax on the difference between sales revenue and total costs plus the tax on other income. Assuming a 35% corporate tax rate for Company A is $0.35\ (25{,}000 - 24{,}000) + 87.50 + 87.50 = \$525$, the tax for Company B is $0.35\ (26{,}000 - 21{,}000) + 87.50$.

and the life of the assets. The effect of differences in assumptions concerning the return on the pension assets depends on the size of the pension assets relative to the size of the firm's earnings. Thus, in comparing the earnings across firms, individual adjustments are necessary if they are to be put on a comparable basis.

The fact that different accounting methods can lead to different reported earnings, together with the belief held by many accountants and managers that earnings are important to the valuation process, has led to another problem. Accountants and management may attempt to manage the level and growth of earnings. There are a number of studies that have examined whether investors can see through attempts to manage earnings. While these studies support the hypothesis that they can, many firms believe the opposite strongly enough that they continue to incur costs in an attempt to manage reported earnings.

In the next part of this chapter we show that despite the problems with accounting earnings, they still represent one of the important inputs in judging a firm's value.

THE IMPORTANCE OF EARNINGS

Several studies have shown that knowledge about past and future earnings can lead to investors earning superior returns despite ambiguity in measuring earnings.

Francis and Schipper (1999) examine the payoff from perfect foreknowledge of the coming year's earnings. In one test, they form two portfolios, one with a positive change in earnings relative to the prior year and the other with a negative change. They form a

hedge portfolio by taking a long position in the positive earnings change firms and a short position in the negative earnings change firms. Over the 1952–1994 sample period, the average 15-month hedge portfolio return is 13.9%.[2] In a second test, they consider both the sign and the magnitude of the coming year's earnings. They form a hedge portfolio by taking a long position in the top 40% of firms ranked by the change in earnings (deflated by the firm's beginning of year market value of equity) and a short position in the bottom 40%. Over the 1952–1994 sample period, the average 15-month hedge portfolio return is 19.6%. By contrast, a hedge portfolio formed by a long position in the top 40% of firms based on the change in cash flows (deflated by the firm's beginning of year market value of equity) and a short position in the bottom 40% earns an average 15-month return of only 6.0% over the same period. These results show that the ability to predict earnings leads to significantly positive returns.

Kormendi and Lipe (1987) examined the effect of earnings information on stock returns. They introduce the notion of earnings persistence and begin to explore the notion of unexpected earnings. Using annual earnings and returns for a sample of 145 firms over the 1947–1980 period, Kormendi and Lipe estimated a regression of the change in earnings against the previous two years' changes in earnings for each firm.[3] The residual from the model represents the firm's unexpected earnings change for the year (i.e., the new information about earnings). From the two autoregressive coefficients, Kormendi and Lipe construct a measure of earnings persistence. Persistence captures the permanence of an earnings change, that is, how much the earnings change continues to the future. The greater the coefficients on the two lagged earnings changes, the greater the persistence, as the lagged changes are more informative about future earnings. The more persistent is an earnings change, the greater is its impact on the entire future earnings series, and thus the greater its stock price impact should be. Kormendi and Lipe estimated the stock market's response to a firm's annual earnings news as the coefficient in a regression of the firm's annual stock return against the firm's unexpected annual earnings change (the residual from the autoregressive model). This coefficient is often referred to as the earnings response coefficient, or ERC. Kormendi and Lipe hypothesized and found that the ERC is positively correlated with earnings persistence; that is, the stock market responds more to the earnings news of firms whose earnings changes are more persistent. This indicates that the market understands the time series properties of a firm's earnings, and stock prices adjust accordingly to earnings information.

Easton and Zmijewski (1989) and Collins and Kothari (1989) extend Kormendi and Lipe's analysis. Easton and Zmijewski show that positive association between the ERC and earnings persistence also extends to the stock market's response to earnings news in the two-day window around the announcement of quarterly earnings. Using an annual return window like Kormendi and Lipe, Collins and Kothari show that the ERC is also positively correlated with firm growth (since greater growth leads to greater future earnings) and negatively correlated with a firm's risk (beta) and market interest rates (since higher risk and interest rates mean a greater discounting of future earnings). Elton, Gruber, and Gultekin (1978) not only looked at the risk-adjusted excess return that could be earned by purchasing stocks on the basis of earnings but also examined the role of forecast data on risk-adjusted excess returns.

---

[2]The 15-month period includes the firm's fiscal year plus the subsequent 3 months. The 3 months are added since a firm's annual report must be released with 3 months after the fiscal year end.

[3]This is known as a second-order autoregressive model. Requiring 34 annual observations for each firm restricted the sample to 145 firms.

The first question they examined was, Do earnings affect prices? If reported earnings are important, then buying those stocks that will experience the largest growth in earnings should lead to an excess risk-adjusted return. To study this question, Elton, Gruber, and Gultekin divided stocks into deciles by the size of the next year's growth in earnings. Then they examined the excess risk-adjusted return that would be earned if each decile were purchased and held until after actual earnings were announced.[4] Stocks that had the highest future growth in earnings provided the highest excess return. The results were statistically significant at the 1% level. Furthermore, the results seem to be economically significant. For example, the 30% of firms that had the highest growth provided an excess risk-adjusted return of 7.48%, while the 30% of firms with the lowest growth (candidates for short sale) provided an excess risk-adjusted return of −4.93%. This provides strong evidence that reported earnings, despite their deficiencies, do impact stock prices.

The next logical subject to look at is the impact of expectational data on stock prices. Economists believe that expectations determine stock prices. If this is true, and the market is efficient, then expectations about future earnings should be incorporated into stock prices. It follows logically that the investor should not be able to make an excess return by either buying or selling stock on the basis of the average (consensus) expectations about future earnings.[5] On the other hand, if prices reflect the consensus estimate, then the investor should be able to earn large excess returns by acting on either the difference between consensus estimates and realizations or changes in the consensus estimates.

Elton, Gruber, and Gultekin examined whether an investor could make an excess return by buying and selling stocks on the basis of the consensus estimate of earnings growth. They divided stocks into deciles based on the consensus forecast of earnings growth. They found that there was no difference in excess return between the deciles. The investor who bought the stocks that were expected to have low growth would have done just as well as the one who bought the stocks with high expected growth. This is what one would expect if markets are reasonably efficient and expectations are reflected in security prices. Their second test involved dividing firms into deciles by the error in the forecast of earnings growth. Here the results were dramatically different. The firms for which the actual earnings growth was higher than the forecasted earnings growth had returns well above normal. The firms with actual earnings growth below estimated earnings growth had returns well below normal. An investor who could forecast earnings better than average could earn excess returns. Finally, Elton, Gruber, and Gultekin divided firms into deciles by the change in the forecast of earnings. This led to even higher returns. While it was profitable to forecast earnings, it was even more profitable to forecast the change in expectations about future earnings. A number of mutual funds have a strategy of buying high-growth firms. By itself, this should not be a useful strategy. What is important is to find high-growth firms that the market believes will be low-growth firms. Even more valuable would be to forecast changes in the market's belief about the future growth of a firm.

The studies just discussed provide strong evidence that earnings affect returns and that superior forecasts of earnings can lead to excess returns. The question is, how much better does the analyst have to be in order to earn excess returns? Table 19.2, taken from the Elton, Gruber, and Gultekin study, provides a partial answer to this question. The table shows the excess return that can be earned if analysts are able to identify the firms whose

[4]To define excess risk-adjusted return, Elton, Gruber, and Gultekin used the methodology outlined in Chapter 17. Each portfolio has its return adjusted by subtracting from actual returns expected returns based on the market model $R_j = \alpha_i + \beta_i R_m$.

[5]The consensus estimate was defined as the average estimate of security analysts at major brokerage houses following a stock. Only stocks followed by three or more analysts were included in the study.

Table 19.2 Excess Returns by Eliminating from Portfolio Those Firms That Had Earnings Estimates the Most Above (or Least Below) Realizations

Percentage of Firms Eliminated	Excess Return If Completely Accurate	Excess Return If 50% Error	Excess Return If 90% Error
0%	0	0	0
10%	1.56	0.78	0.16
20%	2.88	1.44	0.29
30%	3.07	1.53	0.31
40%	4.32	2.16	0.43
50%	5.77	2.88	0.58
60%	7.35	3.67	0.74
70%	9.08	4.54	0.91
80%	9.90	4.95	0.99
90%	10.42	5.21	1.04

Source: Elton, Gruber, and Gultekin (1981).

earnings will be less than the consensus forecast. For example, the second entry in the second column is 1.56%. If the analysts were able to eliminate the 10% of the stocks with the largest overestimate of growth, they would earn 1.56% more than normal, given the risk of the stocks. Similarly, if they were able to eliminate the 20% of stocks with the greatest overestimate of actual growth, an extra 2.88% return above normal would be earned. Columns 3 and 4 show the excess return if there is error in the analysts' ability to select firms with inaccuracies in the average estimate of earnings. The second column assumes that 50% of the time the analyst picks stocks in the category shown (e.g., the 70% of the stocks with the least overestimate of average growth) and 50% of the time she picks stocks that have the average characteristics of the population of stocks. Column 4 is similar, except that it is assumed that the analyst can select from the best category only 10% of the time. As can be seen by examining these columns, even information with little accuracy can lead to excess returns.

This section illustrates the importance of earnings to the valuation process and the importance of being able to forecast earnings. In the next section we examine some time series characteristics of earnings and some methods for forecasting it.

CHARACTERISTICS OF EARNINGS AND EARNINGS FORECASTS

In this section we analyze the characteristics of earnings. Are earnings changes highly related to the performance of the economy? Are future changes in earnings highly related to past earnings? Can analysts forecast earnings? In the last section we saw that good forecasts of earnings can lead to profitable returns. Hence it is important to understand the characteristics of earnings and earnings changes.

The Influence of the Economy and Industry

In earlier chapters we showed that a stock's returns are strongly affected by market movements and by industry or sector returns. A similar phenomenon exists with respect to earnings. Earnings of a firm are strongly influenced by changes in aggregate earnings for the economy, and there is some evidence that they are influenced by changes in the earnings of the industry to which the firm belongs. Table 19.3 illustrates the strength of these influences.

Table 19.3 Proportion of Earnings Movement Attributable to Economy or Industry Influences

Industry	Economy Influence (%)	Industry Influence (%)
Aircraft	11	5
Autos	48	11
Beer	11	7
Cement	6	32
Chemical	41	8
Cosmetics	5	6
Department stores	30	37
Drugs	14	7
Electricals	24	8
Food	10	10
Machinery	19	16
Nonferrous metals	26	25
Office machinery	14	6
Oil	13	49
Paper	27	28
Rubber	26	48
Steel	32	21
Supermarkets	6	33
Textiles and clothing	25	29
Tobacco	8	19
All companies	21	21

Source: Brealey (1969).

The sample used in calculating this table was the earnings from 217 firms for the years 1948–1966. The earnings of the companies that compose the S&P 425 index were used to represent the market index. The companies in the sample were divided into industries and the earnings averaged across each of the companies in the industry to obtain an industry index. The percentage of changes in each firm's earnings that could be attributed to the market and the industry was then determined. The results for individual firms were then averaged across an industry.

As can be seen by examining the table, on average 21% of the changes in firm's earnings can be accounted for by changes in the market's earnings, and an additional 21% of the changes in a firm's earnings can be accounted for by the changes in the industry earnings. The strength of these influences varied considerably. Earnings for companies in industries such as autos, chemicals, and steel seemed to be heavily influenced by market-wide changes. The earnings of firms in the oil industry and the rubber industry seemed to be strongly influenced by industry changes. Many of the differences shown in the table may well be unique to the period examined. However, the large effects of market and industry factors are probably indicative of real influences.[6] A forecast of economy-wide changes and industry-wide changes may be useful first steps in estimating the companies' earnings.

If earnings of firms in the same industry move together, then the announcement of earnings for a given firm should affect not only its own stock price but the stock prices of other firms in its industry. This is referred to as "intra-industry information transfer." Foster (1981) showed that this is indeed the case. Moreover, Foster found that the magnitude of

[6]There is some overstatement of the correlation since the firms themselves are part of the industry and economy.

the information transfer (i.e., stock price impact on nonannouncing firms) is greater for firms that have a larger percentage of their revenues in the same line of business as the announcing firm. Han, Wild, and Ramesh (1989) showed that intraindustry information transfer applied to management forecasts as well: voluntary disclosure of management earnings forecasts affects not only stock prices of the announcing firm but also stock prices of firms in the same industry.

All of these studies strongly suggest that changes in the economy's earnings influence the earnings of many firms. Furthermore, there is some evidence that industry earnings are also important.

To be able to utilize relationships such as those described previously, it is necessary that the relationships be reasonably stable over time and that economy and industry earnings be more easily forecasted than the earnings for individual companies. There is no evidence concerning this. Thus, at this time, although we can say that economy-wide and industry-wide earnings are useful in explaining the earnings of individual companies, demonstration that this is useful in improving prediction must await further research.

Past Earnings and Future Earnings

Two separate issues have been examined with respect to the time series behavior of earnings. One is whether past growth is an indication of future growth. The second is whether the concept of normal earnings is meaningful. We discuss each of these in turn.

One of the popular terms used in the financial literature is the term *growth stock*. This term often refers to a stock that has had substantial growth in the past and is expected to in the future. Names like Apple and Microsoft come to mind. From this, one would expect that stocks that have had high growth in the past would have high growth in the future. A number of studies have seriously questioned this assumption. Lintner and Glauber (1969) examined the correlation of aggregate earnings and earnings per share for 323 companies during 1946–1965. The 20 years of data were divided in four 5-year periods and two 10-year periods. Growth was estimated for each of these periods, and correlations between the growth rates in adjacent periods were calculated.[7] The results give little comfort to anyone expecting past growth to predict the future. The highest association between successive growth rates implied that less than 2% of the variation in growth in the latter period was explained by growth in the earlier period. Lintner and Glauber introduced two modifications to try to improve the correlation in growth rates. First, they deflated earnings by a measure of aggregate economic conditions. Second, they divided firms into groups by stability of growth rate and ran correlations within each group. This did lead to improvement. In one time period they were able to explain almost 50% of the variation in future growth rates by past growth rates. However, for most periods and most cases studied, less than 10% of the variation in future growth rates was explained by past growth.

Brealey (1969) analyzed the same question in a slightly different way. He analyzed the growth of 610 industrial companies from 1950 to 1964. Each year he determined the 305 firms with the highest growth and the 305 firms with the lowest. If past growth is helpful in predicting future growth, then one would expect that firms would tend to have long periods when they were in the high-growth group and long periods when they were in the low-growth groups. The alternative is that the odds of being in either group are 50–50, independent of the firm's position in the previous period. Table 19.4 shows the results. The first column indicates the number of years the firms were in the same group. For example, the first entry in the second column is 1,152. This means that 1,152 times, firms were in

[7]Growth was estimated using a logarithmic regression on time.

Table 19.4 Persistence of Growth

Length of Time in Same Group	No. of Consecutive Years of High Growth	No. of Consecutive Years of Low Growth	Expected No. of Consecutive Years of Low or High Growth If Odds Are 50–50 Regardless of Past Performance
1	1,152	1,102	1,068
2	562	590	534
3	266	300	267
4	114	120	133
5	55	63	67
6	24	20	33
7	23	12	17
8	5	6	8
9	3	3	4
10	6	0	2
11	2	0	1
12	1	0	1
13	0	0	0
14	0	0	0

Source: Brealey (1969).

the high-growth group one year and not in that group the next. The second entry, 562, means that 562 times, firms were in the high-growth group two years in a row and in the low-growth group the next year. The second and third columns look very similar to the last. In fact, the odds of long runs are, in general, higher for the last column than for the second or third. The most striking place where the second and third columns have higher odds than the last is for lengths of time 1. But this implies that a good year follows a bad year more than expected by chance, and vice versa.[8] This is the opposite of what one expects if past growth was a good predictor of the future.

Chan Karceski and Lakonishok (2003) reexamined persistence and predictability of a large sample of U.S. stocks for the period 1951–1999. Over horizons of 5 to 10 years, they find virtually no persistence or predictability in earnings growth, even among firms that are widely reputed to have high growth, such as those in high-tech industries. Moreover, both security analysis long-term growth forecasts and market valuation ratios have little forecasting power to predict differences among firms in long-term growth.

[8]There is a potential bias here. One plus growth from t to $t + 1$ is

$$\frac{\text{Earnings } t+1}{\text{Earnings } t}$$

Likewise, one plus growth from $t - 1$ to t is

$$\frac{\text{Earnings } t}{\text{Earnings } t-1}$$

Thus earnings in t appear on both sides of the equation. Under certain circumstances (such as the reversion to the mean process of generating earnings discussed in the next section), this can cause a negative basis and could account for the results just discussed.

A recent series of articles suggests that when earnings are extreme, the concept of reversion to the mean might provide useful information when forecasting earnings changes.

Brooks and Buskmaster (1976) found that for firms whose current period's earnings were "extreme," next year's earnings reverted to a normal level. In related research, Freeman, Ohlson, and Penman (1982) show that book rates of return (defined as earnings divided by beginning of period book value of owners' equity) are mean reverting: firms whose earnings were high (low) relative to their equity showed declines (increases) in next year's earnings. In other words, book value provides a benchmark for a "normal" level of earnings. Put another way, book value of equity helps predict next period's earnings beyond earnings itself. In a more recent study, Fama and French (2000) also found that mean reversion in earnings was faster when a firm's earnings were further from their mean.

These three studies are typical of the results found by a number of authors. These results have led them to speculate that earnings changes might be independent from period to period. The economic argument of why this might be so goes as follows: The economy is highly competitive. The earnings of a company are subject to a large number of uncertainties not under management control. These include strikes, mineral discoveries, regulatory changes, foreign competition, changing tastes, and so on. These kinds of uncertainties are the dominant influences on a company's fortunes on a year-to-year basis.

The counterargument is that there are a number of companies with monopoly control of the markets, with patent protection on unique products, or with superior management, and these companies are able to sustain a high level of growth over a long period of time.

The argument for independence in earnings is much less persuasive than the argument for the independence of security prices presented in Chapter 17. Earnings are determined by a physical process, while stock prices are determined by expectations. It is reasonable to assume that changes in expectations cannot be predicted from past data or they would already be incorporated in the expectations. It is a more stringent requirement to assume that past levels of a physical process do not convey information about the future. However, the empirical evidence reviewed earlier is a useful cautionary note to those who would place too heavy a reliance on past earnings to predict the future.

The second major issue concerning the time series of earnings is the concept of normal earnings. To understand this issue, it is easiest to ignore growth for the moment and to assume independence of earnings between time periods. One view of a firm's earnings is as follows: The firm's earnings are on average $1.00, but there is some variation. Table 19.6 shows a possible scenario. If this is the process that describes earnings, then one would expect, if earnings were at an extreme, that they would be closer to the mean the next period. For example, if you observed earnings of $1.20 in one period, you would expect, on average, that they would be less the next period. Extremes followed by observations closer to the mean would tend to introduce a negative correlation in the time series.

The second alternative scenario is illustrated in Table 19.6. The distinctive element of this process is that there is no tendency to revert to some mean level of earnings. The change occurs from the last level of earnings. If earnings are $1.20, then 10% of the time they will increase by 10% to $1.32, 20% of the time they will increase by 5% to $1.26, and so forth. This period's earnings serve as a starting point for the change to next period's earnings. If we observe earnings of $1.20, we are just as likely to have an increase in the earnings as a decrease. With this view of earnings there is no such thing as extreme earnings, and one would not expect the negative correlation discussed earlier.

The issue of which process describes earnings is important. If the first process is a better description, then the starting point for any forecast of future earnings is an estimate of the mean or "normal" earnings. If the second process is more descriptive, then the starting point of any forecast is the latest observed level of earnings.

Table 19.5 Possible Levels of Earnings

Earnings	Odds
1.20	10%
1.10	20%
1.00	40%
0.90	20%
0.80	10%

Throughout the discussion, we assumed zero growth. This just simplified the discussion. The same discussion holds with growth. If a process like that in Table 19.5 is descriptive of earnings patterns, except for the presence of a growth rate, then the starting point for the estimate of earnings using historical earnings is normal earnings plus an estimate of growth. If the second model is more descriptive, then the starting point is this period's earnings plus an estimate of growth. We also assumed independence. If there is positive dependence, then this should mitigate the negative correlation in the first case and impart positive correlation in the second. There are two types of evidence on this issue. The first is correlation in successive earnings changes. Brealey examined this question and found slight negative correlation in the series. Although this is supportive of the concept of normal earnings, it was so small that it is not very strong support. The second type of evidence is forecast evidence. Does this concept of normalized earnings lead to a better forecast of earnings, or does using last period's earnings produce a better forecast?

Elton and Gruber (1972) examined this question and found that allowing smoothing over a longer period of time led to better forecasts than did the simple use of last period's earnings. Ball and Watts (1972), in contrast, found that last periods's earnings worked best. There were two major differences in the studies. First, Elton and Gruber utilized much more complicated forecasting models than Ball and Watts.[9] Second, Ball and Watts required that the same forecasting model be used to forecast the earnings of all firms. Elton and Gruber allowed a different model for each firm and selected the one to use in making comparisons that had provided the most accurate forecasting of earnings in prior periods. For many firms this was, in fact, last period's earnings, but in other cases it was a smoothed value of past earnings. When Elton and Gruber allowed this variation, they achieved improved forecasts. Lieber and Ronen (1975) repeated this for the Ball and Watts sample and found that allowing individual variation led to improved forecasting. Thus, reality probably includes both of the models of firms' earnings discussed in Tables 19.5 and 19.6. For many firms the concept of normal earnings is superior, whereas for other firms, last period's earnings provide a better forecast of next period's earnings.

[9]Elton and Gruber (1972) used an exponential smoothing model with an arithmetic change in growth. Their model is presented here: Let E be earnings, g be growth, and subscripts indicate time periods. Let a and b be constants with a value between 0 and 1, and carets (^) indicate smoothed values. Then

1. Forecast of earnings $= \hat{E}_t + \hat{g}_t$
2. $\hat{E}_t = (\hat{E}_{t-1} + \hat{g}_{t-1}) + a[E_t - (\hat{E}_{t-1} + \hat{g}_{t-1})]$
3. $\hat{g}_t = (\hat{g}_{t-1}) + b[(\hat{E}_t - \hat{E}_{t-1}) - \hat{g}_{t-1}]$

In contrast, the Ball and Watts model (1972) was

1. Forecast of earnings $= \hat{E}_t$
2. $\hat{E}_t = a\hat{E}_{t-1} + (1 - a)E_t$

Table 19.6 Possible Changes in Earnings

Earnings Change	Odds
+10%	10%
+5%	20%
0%	40%
−5%	20%
−10%	10%

The research that has been done on using the time series of past earnings to predict change in future earnings is not very encouraging. The evidence seems to suggest that in many cases the naive model of next year's earnings equals this year's earnings seems to do as well as more sophisticated extrapolations. This should serve as a cautionary note to anyone predicting changes in earnings from the past. The exception might be when earnings are abnormally high or low.

Forecasting Earnings with Additional Types of Historical Data

Firms make available a great deal more information than past levels of earnings per share. Perhaps this information can be used to forecast future levels of earnings per share or future growth in earnings per share. For example, changes in sales or research and development expense or new investment might be related to future earnings. Several studies have examined whether past values of other types of historic data could be used to estimate the relationship and whether this relationship could be used to forecast the future.

Ou and Penman (1989) and Lev and Thiagarajan (1993) show that past accounting data can forecast future earnings. Ou and Penman estimate a logit model that explains the sign of next year's change in earnings using a large set of financial statement items. From an initial set of 68 variables their final models include 16–18 statistically significant predictor variables. The development of the final models is primarily data driven. Ou and Penman show that the models have out-of-sample forecasting power, i.e., the models can predict the sign of an earnings change in the future. Since the model estimates the probability of a positive earnings change, they call the fitted prediction PR. Moreover not only can PR predict the sign of next periods earnings change, it can also predict future stock returns. Specifically, Ou and Penman show that a portfolio that is long in stocks with $Pr > 0.6$ (firms with a relatively high probability of a positive earnings change) and short in stocks with a $Pr < 0.4$ (firms with a relatively low probability of a positive earnings change), earns positive future risk adjusted stock returns over the 36 months subsequent to the portfolio formation date. The combination of both earnings and return predictability implies both that fundamental analysis is useful for forecasting accounting earnings, and that the stock market does not appear to appreciate the forecasting value of the accounting data. The market is not semi-strong form efficient with respect to accounting information.[10]

Lev and Thiagarajan (1993) conduct an analysis similar to Ou and Penman's. However, Lev and Thiagarajan's tests are not data driven, but are motivated by an

[10]It is important to point out that Ou and Penman's evidence of market efficiency may be fragile. Holthausen and Larcker (1992) found that Ou and Penman's Pr return predictability strategy did not work in a subsequent period. Stober (1992) and Grieg (1992) interpret Pr's returns as a compensation for risk. Stober finds that Pr's risk-adjusted returns continue for six years, suggesting that the returns include compensation for some uncontrolled risk factor, and Grice finds that size may subsume the Pr effect (although Ou and Penman controlled for size).

analysis of "value drivers" that relate to risk, growth, and competitive position. Thus, Lev and Thiagarajan's set of predictor variables is not only much smaller, 12, but all variables are kept in their model regardless of statistical significance. Lev and Thiagarajan show that the 12 variables can distinguish between high and low growth rates in earnings up to three years ahead.

While the prior studies used the same forecasting model for all firms, Elton and Gruber (1972) explore how using a different model of accounting variables for different types of firms might lead to useful forecasts. They divided firms into groups by similarity in the pattern of their previous growth. They argued that if firms had similar growth patterns, they probably had responded to similar influences. Their procedure yielded a set of 10 groups or pseudo-industries. For each group they estimated a model relating earnings to accounting variables.

This yielded a set of 10 forecast equations, one for each group. When they examined the accuracy of the forecasts generated in this way with the accuracy of a model utilizing only past earnings, they found that the forecasting equations utilizing other firm variables were superior. They repeated the analysis over several periods and several samples, and the results were similar.

The research discussed suggests that firm information, other than past earnings, may be useful in predicting future earnings.

Analysts Forecasts

We have seen in earlier sections that accurate forecasts of earnings can lead to superior returns. It is not surprising that analysts spend a great deal of time and effort forecasting earnings. Given the importance of earnings and the amount of data an analyst possesses, it is not surprising that there is a great deal of research on the accuracy and price impact of analysts' estimates.

There is a large literature on analysts' earnings forecasts, owing to the availability of machine readable data. Brown and Rozeff (1978) were the first to show that analysts' forecasts are more accurate than forecasts based on past earnings time series. While analysts' superiority is due somewhat to their timing advantage (because their forecasts are made after the firm's earnings have been announced and thus are based on more recent information), Brown et al. (1987a,b) show that analysts' forecasting superiority remains even after controlling for timing. This is not surprising, because analysts' forecasts are based on more information than just the past history of earnings. Consistent with this forecasting superiority, Fried and Givoly (1982) show that stock returns at earnings announcements are more highly correlated with earnings surprise (actual earnings minus forecasted earnings) based on analysts' forecasts than earnings surprise based on time series forecasts, indicating that analysts' forecasts are a better representation of the market's unobservable earnings forecast.

Despite their benefits, analysts' forecasts have their drawbacks. Numerous studies, such as Easterwood and Nutt (1999), have found that analysts' forecasts are optimistic (i.e., upwardly biased, compared to actual earnings) and inefficient (failing to incorporate information in past stock price changes or in the analysts' own past forecast errors).[11] While the great majority of the research on analyst forecasts deals with forecasts of quarterly or annual earnings, LaPorta (1996) also finds that analysts' forecasts of long-term earnings growth are also biased. In particular, he finds that firms with the highest (lowest) forecasted growth fall short of (exceed) the estimated growth.

[11]Brown (1998) finds that the bias has declined over time.

Understanding the properties of analysts' forecasts is important, because much of the financial literature uses consensus forecasts as the basis of modeling stock prices. Thus, if analysts are biased, and if their biases are incorporated into stock prices, stock prices can be systematically wrong, that is, inefficient. Because the market responds to analyst forecast errors, investors with advance knowledge of the errors can earn abnormal profits. Put another way, being able to out-forecast the analysts may be rewarding.

LaPorta (1996) and Dechow and Sloan (1997) investigate whether forecast biases lead to return predictability. Both papers attempt to determine whether Lakonishok, Shleifer, and Vishny's (1994) findings that financial ratios (ratios of price to a fundamental signal, such as earnings or cash flows) predict stock returns are due to risk or market inefficiency.[12]

LaPorta sorts stocks on the basis of analyst consensus 5-year earnings growth forecasts. If the market is efficient, then one should not be able to earn excess returns based on known information, such as the five-year growth rate forecast. However, if the market is inefficient because prices are based on the biased forecasts, then excess returns can be earned. He finds that the one-year postformation size-adjusted returns of the lowest forecast decile exceed the one-year returns of the highest forecast decile by an average of 20% over the 1982–1990 period. During the postportfolio formation year, analysts revise their expectations down (up) for the high (low) expectation decile, consistent with the hypothesis that the original expectations were biased. Also during this year, for the high expected growth portfolio, the cumulative returns around the four earnings announcements (3 days centered at each announcement, for a total of 12 days) are −1.6%, suggesting that the returns are not due to risk (unless this portfolio is a risk hedge). Finally, also inconsistent with the risk hypothesis, the low expected growth portfolio does not have a higher beta or return standard deviation than the high expected growth portfolio. Consistent with LaPorta's findings, Dechow and Sloan find that stock prices naively incorporate analysts' long-term earnings growth forecasts; actual earnings grow at less than half the rate forecast by analysts, but stocks initially reflect essentially all of the forecasted earnings growth. In summary, both LaPorta's and Dechow and Sloan's evidence is consistent with market inefficiency with respect to forecasted long-term earnings growth, and stock prices seem to reflect the biased growth rates. This evidence is consistent with studies cited earlier that being able to forecast changes in the one-year consensus estimate of earnings leads to a higher return than forecasting actual earnings.

In Chapter 26 we return to the question of the accuracy of analysts' forecasts. We place special emphasis on techniques for determining the accuracy of these forecasts.

CONCLUSION

In Chapter 26, we analyze the accuracy of analysts' estimates in some detail. The studies are mixed as to the predictive content of these estimates and their impact on prices.

The studies discussed in this section do not provide a magic formula for predicting earnings. This should not be surprising or especially disturbing. Even if such a formula existed, its value would already be mitigated as investors utilized it to obtain superior predictions, and this was reflected in security price. We view the studies discussed in this section as suggestive of the kinds of analysis that might be worthwhile as well as the types of behavior and research that are unlikely to be productive. Research is under way and should continue in this area.

[12]Lakonishok, Shleifer, and Vishny (1994) suggest that the predictable returns are due to investors naively extrapolating past growth. They refer to this as the *naive expectations hypothesis*.

QUESTIONS AND PROBLEMS

1. Write down the forecast of next period's earnings if

A. Earnings are a mean reverting process with no trend or cycle.

B. Earnings are a mean reverting process with a trend but not a cycle.

C. Earnings are a mean reverting process with a trend and a cycle.

2. How would earnings be forecast if there was a strong relationship between the firm's earnings and the industry's and economy's earnings?

3. Is a strong relationship between a firm's earnings and an economy's earnings consistent with a mean reversion process for earnings generation?

4. Is a strong relationship between a firm's earnings and an economy's earnings consistent with last period's earnings being a better estimate of next period's earnings than normal earnings?

5. If expectations determine share price, what is a valuable analyst?

BIBLIOGRAPHY

1. Ball, Ray, and Watts, Ross. "Some Time Series Properties of Accounting Numbers," *Journal of Finance*, **27** (June 1972), pp. 663–681.
2. Bar-Yosef, Sasson, Callan, Jeffrey R., and Livnot, Joshua. "Causality and Autoregressive Modeling of Earnings-Investment," *Journal of Finance*, **42,** No. 1 (March 1987), pp. 11–28.
3. Brealey, Richard. *An Introduction to Risk and Return from Common Stocks* (Cambridge, MA: MIT Press, 1969).
4. Brooks, Leroy, and Buckmaster, Dale. "Further Evidence of the Time-Series Properties of Accounting Income," *Journal of Finance*, **XXXI,** No. 5 (1976), pp. 1359–1373.
5. Brown, Lawrence, and Rozeff, Michael. "The Superiority of Analyst Forecasts as Measures of Expectations: Evidence from Earnings," *Journal of Finance*, **XXXIII,** No. 1 (March 1978), pp. 1–16.
6. Brown, Lawrence. "Analyst Forecast Errors: Additional Evidence," *Financial Analysts Journal*, **53 (**1997), pp. 81–88.
7. Brown, Lawrence, and Rozeff, Michael. "The Superiority of Analysts' Forecasts as Measures of Expectations: Evidence from Earnings," *Journal of Finance*, **33** (1978), pp. 1–16.
8. Brown, Lawrence, Griffin, Paul, Hagerman, Robert, and Zmijewski, Mark. "Security Analyst Superiority Relative to Time-Series Models in Forecasting Quarterly Earnings," *Journal of Accounting and Economics*, **9** (1987a), pp. 61–87.
9. ——. "An Evaluation of Alternative Proxies for the Market's Expectation of Earnings," *Journal of Accounting and Economics*, **9** (1987b), pp. 159–193.
10. Chan, Louis, Karceski, Jason, and Lakonishok, Josef. "The Level and Persistence of Growth Rates," *Journal of Finance*, **LVIII,** No. 2 (2003), pp. 643–684.
11. Chant, Peter D. "On the Predictability of Corporate Earnings per Share Behavior," *Journal of Finance*, **35,** No. 1 (March 1980), pp. 13–22.
12. Collins, Daniel W., and Kothari, S. P. "An Analysis of Intertemporal and Cross-Sectional Determinants of Earnings Response Coefficients," *Journal of Accounting and Economics*, **11,** No. 2–3 (July 1989), pp. 143–181.
13. Copelåd, Ronald, and Marioni, Robert. "Executives' Forecasts of Earnings per Share versus Forecasts of Naive Models," *Journal of Business*, **45,** No. 4 (Oct. 1972), pp. 497–512.
14. Cragg, J. G., and Malkiel, Burton. "The Consensus and Accuracy of Some Predictions of the Growth of Corporate Earnings," *Journal of Finance*, **XXIII,** No. 1 (March 1968), pp. 67–84.
15. Deehow, Patricia, and Sloan, Richard. "Returns to Contrarian Investment Strategies: Tests of Naïve Expectations Hypothesis," *Journal of Financial Economics*, **43** (1997), pp. 3–27.

16. Deschamps, Benôit, and Mehta, Dileep R. "Predictive Ability and Descriptive Validity of Earnings Forecasting Models," *Journal of Finance*, **35,** No. 4 (Sept. 1980), pp. 933–950.
17. Easterwood, John, and Nutt, Stacey. "Inefficiency in Analysts' Earnings Forecasts: Systematic Misreaction or Systematic Optimism," *Journal of Finance*, **54** (1999), pp. 1777–1797.
18. Easton, Peter, and Zmijewski, Mark. "Cross-Sectional Variation in the Stock Market Response to Accounting Earnings Announcements," *Journal of Accounting and Economics*, **11** (1989), pp. 117–141.
19. Edwards, Charles, and Hilton, James. "Some Comments on Short-Run Earnings Fluctuation Bias," *Journal of Financial and Quantitative Analysis*, **V,** No. 2 (May 1970), pp. 187–201.
20. Elton, Edwin J., and Gruber, Martin. "Improved Forecasting through the Design of Homogeneous Groups," *Journal of Business*, **44,** No. 4 (Oct. 1971), pp. 432–450.
21. ——. "Earnings Estimation and the Accuracy of Expectational Data," *Management Science*, **18,** No. 2 (April 1972), pp. 409–424.
22. Elton, Edwin, Gruber, Martin, and Gultekin, M. "The Usefulness of Analyst Estimates of Earnings," unpublished manuscript (1978).
23. ——. "Expectations and Share Prices," *Management Science*, **27,** No. 9 (Sept. 1981), pp. 975–987.
24. ——. "Professional Expectations: Accuracy and Diagnosis of Errors," *Journal of Financial and Quantitative Analysis*, **19,** No. 4 (Dec. 1984), pp. 351–364.
25. Fama, Eugene, and French, Kenneth. "Forecasting Profitability and Earnings," *Journal of Business*, **73,** No. 2 (2000), pp. 161–175.
26. Foster, George. "Intra-Industry Information Transfers Associated with Earnings Releases," *Journal of Accounting and Economics*, **3 (**1981), pp. 201–232.
27. Francis, Jennifer, and Schipper, Catherine. "Have Financial Statements Lost Their Relevance?" *Journal of Accounting Research*, **37,** No. 2 (1999), pp. 319–352.
28. Freeman, Robert, Ohlson, James, and Penman, Stephen. "Book Rate-of-Return and Prediction of Earnings Changes: An Empirical Investigation," *Journal of Accounting Research*, **20,** No. 2 (1982), pp. 639–653.
29. Fried, Dov, and Givoly, Dan. "Financial Analysts' Forecasts of Earnings: A Better Surrogate for Market Expectations," *Journal of Accounting and Economics* **4** (1982), pp. 85–107.
30. Gonedes, Nicholas. "Evidence on the Information Content of Accounting Numbers: Accounting-Based and Market-Based Estimates of Systematic Risk," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 3 (June 1973), pp. 407–443.
31. ——. "A Note on Accounting-Based and Market-Based Estimates of Systematic Risk," *Journal of Financial and Quantitative Analysis*, **X,** No. 2 (June 1975), pp. 355–367.
32. Grieg, Anthony. "Fundamental Analysis and Subsequent Stock Returns," *Journal of Accounting and Economics*, **15 (**1992), pp. 413–442.
33. Han, Jerry, Wild, John, and Ramesh, K. "Managers' Earnings Forecasts and Intra-Industry Information Transfers," *Journal of Accounting and Economics*, **11** (1989), pp. 3–33.
34. Holthausen, Robert, and Larcker, David. "The Prediction of Stock Returns Using Financial Statement Information," *Journal of Accounting and Economics*, **15** (1992), pp. 373–411.
35. Kormendi, Roger, and Lipe, Robert. "Earnings Innovations, Earnings Persistence, and Stock Returns," *Journal of Business*, **60,** No. 3 (1987), pp. 323–345.
36. Lakonishok, Josef, Shleifer, Andrei, and Vishny, Robert. "Contrarian Investment, Extrapolation, and Risk," *Journal of Finance*, **49** (1994), pp. 1541–1578.
37. LaPorta, Rafael. "Expectations and the Cross-Section of Stock Returns," *Journal of Finance*, **51** (1996), pp. 1715–1742.
38. Lev, Baruch, and Thiagarajan, Ramu. "Fundamental Information Analysis," *Journal of Accounting Research*, **31,** No. 2 (1993), pp. 190–215.
39. Lieber, Zvi, and Ronen, Joshua. "Earnings Estimates and Historical Data," unpublished manuscript, Ross Center, New York University (1975).
40. Lintner, John, and Glauber, Robert. "Higgledy-Piggledy Growth in America," unpublished manuscript (1969).
41. Lorie, J., and Hamilton, M. *The Stock Market: Theories and Evidence* (Homewood, IL: Richard D. Irwin, 1973).

42. Mastrapasqua, Frank, and Bolten, Steven. "A Note on Financial Analyst Evaluation," *Journal of Finance*, **XXVIII,** No. 3 (June 1973), pp. 707–712.
43. McEnally, Richard. "An Investigation of the Extrapolative Determinants of Short-Run Earnings Expectations," *Journal of Financial and Quantitative Analysis*, **VI,** No. 2 (March 1971), pp. 687–706.
44. Newell, Gale. "Revisions of Reported Quarterly Earnings," *Journal of Business*, **44,** No. 3 (July 1971), pp. 282–285.
45. Niederhoffer, V., and Regan, P. "Earnings Changes, Analysts' Forecasts, and Stock Prices," *Financial Analysts Journal*, **28,** No. 3 (May–June 1972), pp. 65–71.
46. Ou, Jane, and Penman, Stephen. "Financial Statement Analysis and the Prediction of Stock Returns," *Journal of Accounting and Economics*, **11** (1989), pp. 295–329.
47. Penman, Stephen H. "The Predictive Content of Earnings Forecasts and Dividends," *Journal of Finance*, **38,** No. 4 (Sept. 1983), pp. 1181–1200.
48. ——. "A Comparison of the Information Content of Insider Trading and Management Earnings Forecasts," *Journal of Financial and Quantitative Analysis*, **20,** No. 1 (March 1985), pp. 1–18.
49. Richards, Malcolm. "Analysts' Performance and the Accuracy of Corporate Earnings Forecasts," *Journal of Business*, **49,** No. 3 (July 1976), pp. 350–357.
50. Stober, Thomas. "Summary Financial Statement Measures and Analysts' Forecasts of Earning," *Journal of Accounting and Economics*, **15 (**1992), pp. 347–372.

20

Behavioral Finance, Investor Decision Making, and Asset Prices

> Even apart from the instability due to speculation, there is the instability due to the characteristic of human nature that a large proportion of our positive activities depend on spontaneous optimism rather than mathematical expectations, whether moral or hedonistic or economic.
>
> —John Maynard Keynes, *General Theory of Employment, Interest, and Money* (1936)

Most of the chapters in this book are normative—that is, they are concerned with how investors should make choices. In practice, however, many people make suboptimal economic or financial decisions. For many reasons it is useful to understand how and why this happens. First, and most importantly, such knowledge can help to improve future decision making. If there are a few basic mistakes that investors make repeatedly, it may be possible through education, training, and communication to reduce or eliminate these tendencies. Second, to the extent that certain forms of behavior are pervasive in the market, they may influence security prices. In the first section of this chapter, we discuss the theory and evidence about investor psychology and behavior and the possibility that these phenomena may play a role in investor decision making. In the second section, we examine whether investor psychology actually influences asset prices.

PROSPECT THEORY AND DECISION MAKING UNDER UNCERTAINTY

The central challenge to investors is the problem of decision making under uncertainty. A major area of research in finance is the positive question of how people *actually* make decisions when faced with risk. For example, a common puzzle observed by economists is why people buy lottery tickets when the expected value of such an "investment" is less than the cost of the ticket—behavior that is inconsistent with most common utility functions. Harry Markowitz (1952) proposed one of the earliest solutions to this problem by suggesting that investor attitudes about gambles of different amounts were implicitly compared to their "customary wealth," and gambles for large amounts compared to customary

wealth are treated more conservatively. In other words, a willingness to gamble depended very much on the status quo.

Markowitz's model implies that the utility function of an investor is convex in some places and concave in others—quite different from the typical assumption of everywhere concave utility. While classical financial models generally consider a consistent posture of risk aversion to be a rational attitude toward uncertainty, a number of researchers since Harry Markowitz have explored the evidence for models in which risk aversion (or risk seeking) depends very much on the way the risks are framed and conceptualized by the investor. In these models, investor psychology, mood, and mental "shortcuts" or heuristics play a large role in determining investor choice.

An Experiment

The following questions were put to a set of participants in a psychological study. The percentage of responses is given in brackets after each question:

> Imagine that the United States is preparing for the outbreak of an unusual Asian disease that is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimate of the consequences of the programs are as follows:
>
> - If Program A is adopted, 200 people will be saved. [72% chose this]
> - If Program B is adopted, there is a one-third probability that 600 people will be saved and a two-thirds probability that no people will be saved. [28% chose this]
>
> Which of the two programs do you favor?

Another group was faced with the same problem but a different way of expressing the probabilities:

- If Program C is adopted, 400 people will die. [22% chose this]
- If Program D is adopted, there is a 1/3 probability that nobody will die, and 2/3 probability that 600 people will die. [78% chose this]

Which of the two programs do you favor?

Notice that Programs A and C are identical and B and D are identical. However, subjects responded quite differently to the idea of gambling to "save" lives versus gambling on the loss of life. They regarded potential gains and losses differently, and it affected their decisions about treatment—even though there is no objective difference between the two.

This experiment was conducted by Daniel Kahneman and the late Amos Tversky (1979), two of the leading figures in the study of investor psychology and choice. Based on numerous experiments such as this one, they developed a model of investor decision making under uncertainty called Prospect Theory. Prospect Theory seeks to explain decisions that are inconsistent with rational probability assessments and standard utility functions. Like the early Markowitz utility model, Prospect Theory posits an asymmetric attitude toward risk, depending on how the potential gains or losses relate to a certain reference point. This reference point could be current wealth, a neighbor's wealth, or the price at which an asset is purchased. Kahneman and Tversky's utility function is concave above the given reference point and convex below it. This structure creates risk aversion with respect to gains and risk seeking with respect to losses and can lead to different decisions depending on whether the outcomes are posed as gains or losses.

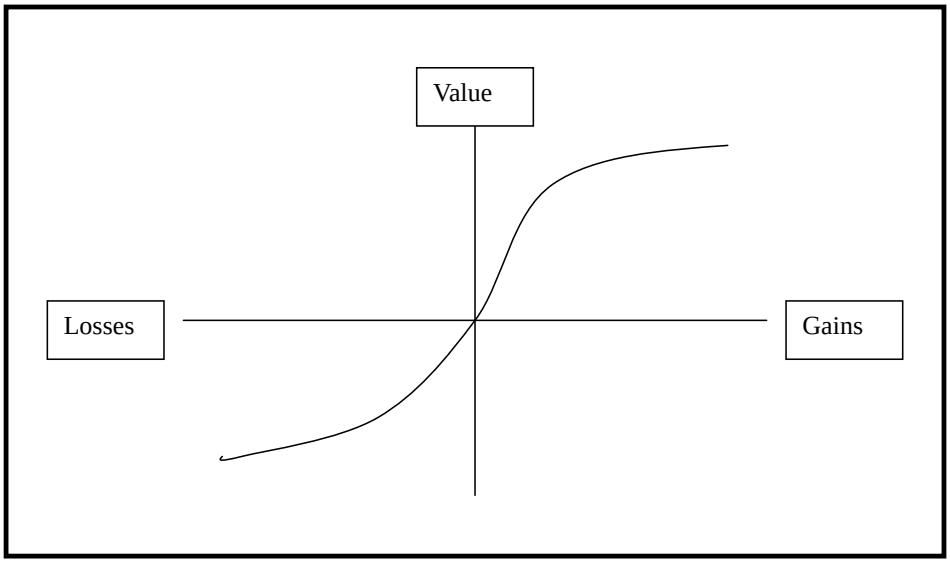


Figure 20.1 Prospect Theory utility function concave in gains and convex in losses.

Figure 20.1 shows the prospect theory value function as proposed by Kahneman and Tversky.

Note that it has a specific reference point—in this figure that point is the intersection for which there are neither gains nor losses. The shift from risk loving to risk averse is indicated by an inflection at this point. Below it, investors will be motivated to lock in gains but not to realize losses. Thus, when they face a potential loss, they prefer to gamble. Tests of Prospect Theory have focused on this asymmetric attitude toward risk.

The Disposition Effect

Consider the case in which an investor bought a stock a month ago for $50 and its current price is $40. Suppose that the chances of the stock going up or down by $10 in the next month are 50/50. An investor with a Prospect Theory valuation function faces the choice either to sell the stock now and recognize a $10 loss or hold the stock for another month. Because the stock has gone down, the investor will be more prone to taking the risk of a 50/50 gamble next period to "make up" the paper loss.[1]

Hirsh Shefrin and Meir Statman (1985) term this attitude the *disposition effect* and point out that it is contrary to good tax planning—in other words, there are negative economic consequences to realizing capital gains and deferring capital losses. They also find circumstantial empirical evidence that investors tend to "ride" their losses throughout the year rather than rationally taking them when a normative model of tax selling would propose.

In a test of the pervasiveness of the disposition effect among investors, Terrance Odean (1998) used a large database of individual brokerage accounts active in the 1990s to show conclusively that the average investor in the sample was more prone to recognizing gains as

[1]This example paraphrased from Shefrin and Statman (1985).

opposed to losses, a finding that can be interpreted as empirical support for the prospect theory model of investor choice. Grinblatt and Keloharju (2001a), using a data set comprising virtually all investors trading in Finland, also document a strong tendency not to recognize losses. The value of this study is that they are able to observe virtually all domestic stock trading in a single country. This allows them to see who is the buyer and seller on each side of the trade in order to determine what investor characteristics are associated with buying stocks versus selling. Dhar and Zhu (2005) discovered that the tendency towards the disposition effect differs significantly across investors. Those with higher levels of education and wealth are less likely to be prone to this asymmetric behavior. In a study far afield from investing, Chen, Santos, and Lakshminarayanan (2005) have identified loss-aversion behavior in Capuchin monkeys, suggesting that loss aversion is instinctive.

BIASES FROM LABORATORY EXPERIMENTS

In their laboratory experiments, Kahneman and Tversky discovered that subjects confronted with uncertainty will typically use mental shortcuts or *heuristics* to guide their decisions. These heuristics may lead to biased or poor choices under uncertainty; however, subjects seemed to consistently rely on them anyway.

Heuristics

Representativeness is a tendency to stereotype a situation through a conceptual analogy to a "representative" type or jump to conclusions based on limited information. Experiments reveal that people often draw inferences about probabilities without considering important issues such as sample size and tend to extrapolate beliefs from an isolated experience.

Anchoring and adjustment is the tendency to "anchor" your understanding of a situation on a familiar one and then to make modest adjustments for the perceived differences. For example, if a stock has unknown expected returns, one might naturally anchor expected return on the S&P 500 and then make an adjustment for risk or for industry.[2]

Availability bases probability assessments on recent or "visible" events rather than the entire range of relevant data, for example, by looking at only recent trends in stock prices to predict future returns.

Overconfidence is the tendency to overestimate one's personal ability to accurately estimate the range of outcomes of a gamble.

These biases, first discovered in laboratory situations, have become the subject of many empirical and theoretical studies in finance. Daniel Kahneman was awarded the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel in 2002 for his joint work with the late Amos Tversky on "bounded" rationality in decision making, of which these heuristics and Prospect Theory are now classical examples.

Other Biases

Several other researchers have identified and tested cognitive models, heuristics, and biases in financial decision making. While these are related to the pathbreaking work of

[2]Although not often thought of as a "behavioral" economist, Irving Fisher was one of the first to argue that a widespread conceptual bias prevented investors from rational portfolio choice. Fisher called this bias "money illusion" and marveled at the fact that people ignored the changes in the purchasing power of the currency, and tended to "anchor" their understanding of economic value to the currency, rather than understanding that the currency itself was subject to inflation.

Kahneman and Tversky, they are additional distinct effects that provide further efforts on the bounded rationality of investors.

Cognitive Dissonance One psychological model that preceded the pathbreaking work of Kahneman and Tversky is Leon Festinger's theory of cognitive dissonance. Festinger's experiments focused on difficult purchase decisions under uncertainty and the cognitive processes that helped to ex post rationalize these decisions. Festinger found, for example, that after subjects made a purchase choice for an appliance, they tended to focus on, and selectively recall, advertisements positively affirming the taken decision and to filter out contradictory evidence. Festinger's model of ex post rationalization through biased perception and recall is based on the notion that hard-to-make decisions—such as the choice between two very similar products—created the future potential for stress between past actions and present conditions. Festinger (1957) theorized that the greater the stress and uncertainty surrounding the decision, the more powerful the ex post filtering and rationalization processes. Akerlof and Dickens (1982) suggest that people with hazardous jobs use cognitive dissonance about the probability of mishap to justify their employment to themselves. Goetzmann and Peles (1997) test the cognitive dissonance hypothesis in a study of the mutual fund holdings and retirement accounts of a number of investors. Hypothesizing that the stress of personally choosing a mutual fund is a potential source of dissonance, the study tested and found that investors systematically overestimated the previous year returns to their mutual funds.

Mental Accounting Mental accounting is the failure to consider all elements of the portfolio as an integrated whole. Optimization theory tells us that making choices over parts of a portfolio without considering it in its entirety will almost certainly lead to suboptimal decisions. Shefrin and Statman (2000) argue that this approach to financial decision making is pervasive. People conceptually (and sometimes actually) place assets in separate "accounts" and treat them differently. This is also sometimes called the "house money effect" because a gambler might be less risk averse with the gains from the evening than with the money won to cover her initial stake. They find that separate mental accounting will lead investors to treat one part of their portfolio like a "nest egg" and another part of the portfolio like a lottery ticket. Massa and Simonov (2003) test for this behavioral bias using a large data set of individual investor accounts in Sweden and find a strong tendency for investors to treat previous year's gains as "house money."

Mood and Emotion Making choices based on mood and emotion as opposed to rational valuation is not a model in the conventional sense; however, mood and emotion (or *affect* in psychological terms) may play an important role in investor decision making. As indicated earlier, Festinger included emotion as an element of the theory of cognitive dissonance. The quote from Lord Keynes at the beginning of this chapter implies that economists have long believed that emotion can play a potentially important role in economic life.

Recently, a number of studies have sought evidence that emotion influences investor choice and perhaps even security prices. Since investor mood is difficult to measure directly, researchers have used environmental factors known to affect mood. For example, Kamstra, Kramer, and Levi (2003) focus on seasonal affective disorder (SAD), which links depression to the amount of winter sunlight. They find some evidence that markets in which the potential for SAD is greatest exhibit significant seasonal variation in returns. They conjecture that this is due to the changing risk aversion of SAD-prone investors. In a similar vein, another set of researchers have documented evidence of secular changes in stock returns associated with lunar cycles. Still others have interpreted

daily fluctuations in stock market returns to weather-related optimism or pessimism. Stock markets in the United States and around the world are slightly more prone to positive returns on days with less cloud cover. Hirshleifer and Shumway (2003) attribute this to investor sentiment, while Goetzmann and Zhu (2005) and Linnainmaa and Rosu (2009) find evidence that it operates though the provision of liquidity.

Recently, financial researchers have attempted to measure affect in real time through documenting physiological changes experienced by investors. Lo and Repin (2002) attach wires to a number of professional traders at a hedge fund and study their response to risk. They find that news and volatility elicit emotional responses and experienced traders remain calmer in these circumstances—in other words, emotion plays a role in the process of trading. However, does it affect decision making? An interdisciplinary study involving a team of neurophysiologists and behavioral economists studied this question by asking a number of cognitively impaired subjects to play an investment game. Subjects with lesions in the emotion centers of the brain made better investment decisions and earned more money than unimpaired players. The natural interpretation is that emotional response to uncertainty prevented rational decision making. Experimental evidence indicates that emotional impairment to judgment might in fact be associated with specific structures in the brain—that is, our investing biases and heuristics might be "hardwired."

These physiological findings have recently stimulated theorists to develop significantly more sophisticated models of cognition that incorporate both rational and emotional mental processes. Anat Bracha (2004a, 2004b) builds a model in which the brain keeps separate "accounts" that interact according to their own, separate goals. Decisions under uncertainty—such as the choice about purchasing insurance—are the result of an equilibrium between these two mental personas. Loewenstein and O'Donoghue (2004) propose a similar but less structured "dual" model of cognition that specifies a role for the emotional part of the brain in decision making. Perhaps these and other future theoretical contributions will provide a framework for understanding what factors can maximize rational decision making and identify situations that are most likely to lead to poor investment choice.

Local Bias Another widely studied pattern of investor behavior is the tendency to invest in the stocks of local companies. Because of the benefits of international diversification, financial researchers have long been aware of, and to a large extent critical of, the tendency of investors to over-weight domestic stocks in their portfolio. Huberman (2001) discovered that this tendency to invest locally extends to U.S.-only portfolios as well. Looking at the ownership of the regional telephone companies in the United States, he found overwhelming evidence that their shares tended to be owned by investors in their own region.

One rational reason for this tendency to invest close to home is that local investors might have more information about nearby companies. Ning Zhu (2005) documents the tendency to trade in local stocks among a large set of individual investors and finds no evidence that these trades produce superior returns. On the other hand, Ivkovitch and Weisbrenner (2003) obtain opposite results with the same data when they focus on investor holdings rather than trades. Results favoring superior information as the basis for local trading are obtained by Massa and Simonov (2003), who use a large database of Swedish investors. Kumar (2005) is able to partially reconcile these ambiguities and largely confirm that the tendency to invest close to home, at least among U.S. investors, is less a function of superior information and more a function of investor confidence in decisions about local companies. Coval and Moskowitz (1999, 2001) document these

same tendencies among institutional investors, although they find some support for better stock picking closer to home by mutual fund managers. Grinblatt and Keloharju (2001) show that Finnish investors are strongly influenced by geography in their choice of companies in which to invest. The propensity to look close to home when deciding where to invest appears pervasive. How much of this local bias is driven by informational advantage remains a lively topic of debate, with fascinating evidence emerging from studies of individual investor accounts from all over the world.

The Path of Least Resistance An extreme passivity in decision making by human participants in pension funds has been observed. Several studies have shown that some elements of employee decisions are consistent with taking the path of least resistance. Choi, Laibson, Madrian, and Metrick (2002) find that participants raise their participation when automatic enrollment is offered in 401(k) plans; the vast majority of participants accept the automatic default investment plan when it is offered. Elton, Gruber, and Blake (2006) find that investors hold more money in company stock when the company makes its contribution in the form of company stock. An excellent overview of the power of default options on investment decision making can be found in Beshears et al. (2009).[2]

Diversification Heuristic Benartzi and Thaler (2001) document that in many choice situations, people tend to take equal amounts of each choice when they would not if the choices were presented sequentially. For example, at the first house they came to, Halloweeners were told to pick two candy bars from a container holding Milky Ways and Three Musketeers. They almost always picked one of each. However, if they were offered the choice of one candy bar from a bowl of Milky Ways and Three Musketeers at two different houses, they almost always picked the same candy at each house. Benartzi and Thaler argue this type of behavior leads investors to place close to an equal amount in their choices in pension plans even when this is not optimum. For example, in their study, they found TWA employees were offered four stock funds and one bond fund in their pension plan, while University of California employees were offered four bond funds and one stock fund. The TWA employees put 75% in stock, while the University of California employees put 34% in stock. They found similar results when they experimentally asked investors their allocation over different sets of options and varied the number of bonds and stock funds in the mix. This and similar studies find that investors' allocation is affected by the number of choices of each type offered. If they are offered three government bond funds in their choice for their pension money, they will place a much greater amount in government bond funds than if they are offered one. This is true even when the amount to place in government bonds is largely unaffected by this difference.

SUMMARY OF INVESTOR BEHAVIOR

In sum, the past two decades of research about investor choice suggests that many investors make suboptimal decisions. People tend to make mistakes in predictable ways that reflect the use of heuristics, or mental shortcuts. Laboratory experiments—most significantly those by Kahneman and Tversky—have been the basis for identifying and labeling these cognitive heuristics and using them to conjecture nonstandard forms of utility functions. Analysis by other investigators has widened our understanding of the limitations of investor rationality. Empirical evidence using data on individual investor decisions has tended to demonstrate the pervasive use of these heuristics—not all of which are consistent with a particular utility function or even with each other. Researchers have documented or conjectured other patterns of behavior associated with decision making under

uncertainty—from cognitive dissonance to home bias to transactions impaired by the emotional centers of the brain.

The strong message from these studies is that investors do not always act rationally and in their own best interest. The evidence in this chapter strongly argues for normative financial research. Indeed, in light of this evidence, some have called for less investor choice over savings and retirement, not more. Should investors, for example, be constrained from investing "too much" in local companies? Is it wiser to delegate your portfolio choice to someone who can be less emotional about it? While these actions may help in certain circumstances, the empirical studies of local investing suggest that this might limit some beneficial, informed trading.

BEHAVIORAL FINANCE AND ASSET PRICING THEORY

The numerous empirical studies of investor choice under uncertainty presented in the first section of this chapter have largely confirmed what most people have long suspected, and what P. T. Barnum is commonly believed to have so eloquently articulated: Documenting investor irrationality is important because it motivates the need for widespread education about finance, and perhaps also, to some extent, for regulatory protection of investors. On the other hand, does investor psychology actually influence asset prices?

To address this question, it is worth reviewing what neoclassical theory does and does not say about price formation. Although the original form of the capital asset pricing model implies a rigid adherence by all investors to precisely the same holdings of risky assets in precisely the same proportion, the Arbitrage Pricing Theory (APT) developed by Stephen A. Ross actually requires very little in terms of rational investor behavior. The APT argues that when a security begins to drift away from the security market line (or plane), the actions of observant and sufficiently capitalized speculators who are willing to accept some risk to achieve a positive return will increase the demand for (or supply of) the nonequilibrium priced security, and the price will be driven back toward the security market plane. The theory relies upon this canny speculator to be the marginal investor; however, it also allows for others to trade. In fact, without the influence of less wise investors, the price would never drift away from the security market line. The APT thus allows for considerable cross-sectional variation in investor skill and reliance upon accurate versus biased valuation models. The APT will only drive prices to fundamental value when there is a liquid and well-developed capital market. This market must be characterized by the opportunity to engage in arbitrage and sufficient financial capital for some investors to do so.

Opportunity

A key requirement of the APT is the ability of at least some speculators to engage systematically in the arbitrage process. For example, commodity funds cannot be sold short, so recognition that they are overpriced does not allow an astute speculator to engage in arbitrage. In fact, a number of researchers studying investor behavior during the later 1990s have pointed out that prices for certain securities far exceeded reasonable economic values—perhaps reflecting investor foolishness. For example, Michael Rashes (2001) finds that a stock with the ticker symbol MCI changed price whenever important news about the telecommunications company MCI was released. The surprising thing about this pattern is that shares in the telecom company MCI trade on the NASDAQ with the symbol MCIC. MCI is the New York Stock Exchange ticker symbol for Massmutual Corporate Investors, a $200 million closed-end bond fund, not a telecommunications stock at all.

Its comovement with telecommunications stocks was due entirely to investors confused by the ticker symbol. Why did arbitrageurs not exploit this irrational behavior? They did not exploit it because the small size and relative illiquidity of Massmutual Corporate Investors made transactions costly and potential arbitrage profits small.

Other historical examples of apparent irrational deviations from economic value appear to share similar barriers to arbitrage. Limits on the ability of speculators to short securities appear to allow for inflated asset values and inefficient pricing.[3] Prices may deviate from fundamental values in situations in which it does not pay arbitrageurs to exploit the spread in prices.

Financing

Financing is nearly as important as opportunity for the arbitrageur. If the arbitrageur needs to borrow to fund his purchases or sales, then even the smartest speculator must face the risk of going bankrupt before prices are driven back to their economically true value. This is the key insight of Shleifer and Vishny (1995), who explicitly model the limits of arbitrage. They point out that, in a world in which credit-constrained speculators cannot always arbitrage away deviations from the SML, investor sentiment itself can become a risk. The old Wall Street adage "Don't fight the tape" captures this basic intuition. Even if the smart investor knows the price of an asset is wrong, there is no way to exploit that knowledge when everyone believes otherwise. While the APT implies that the marginal investor is likely to be a canny speculator, the "limits to arbitrage" model implies that the marginal investor could be the average investor, or part of a large group of investors with a particular conviction about the value of a security—right or wrong. When opportunities to exploit mispricings are limited, or the financing for such activity is constrained, asset prices may reflect the beliefs, emotions, and biases of ordinary investors.

Asset Prices and Demand Curves

Consider a world that matches the assumptions of the APT. In this world, for some reason, an investor needs to sell a large position in one stock. Who will buy the stock and what price will the person pay for it? This situation can be represented by a simple intersection of supply and demand curves, as shown in Figure 20.2.

The horizontal line is the demand curve. It indicates that the price the market will pay for the stock does not depend on the quantity for sale. Because the present value of the shares does not depend on the quantity for sale at a given time, the arbitrageurs in the market stand willing to buy all the available shares for a given price.

Now imagine a situation in which the arbitrageurs have borrowing costs, or limits in their ability to take the other side of the offered trade. This situation is represented by the dashed line. It slopes down, indicating that the marginal investor—the price setter in the market—is unable to buy unlimited quantity of the stock. This is a necessary condition for investor psychology to affect the market—at least if one believes that arbitrageurs are

[3]For example, Owen Lamont and his coauthors have demonstrated that short-sales constraints appear to regularly restrict arbitrageurs from exploiting overpriced securities. See Jones and Lamont (2002) and Lamont and Thaler (2003). Bris, Goetzmann, and Zhu (2004) find that short-sales constraints in international markets are associated with inefficient pricing.

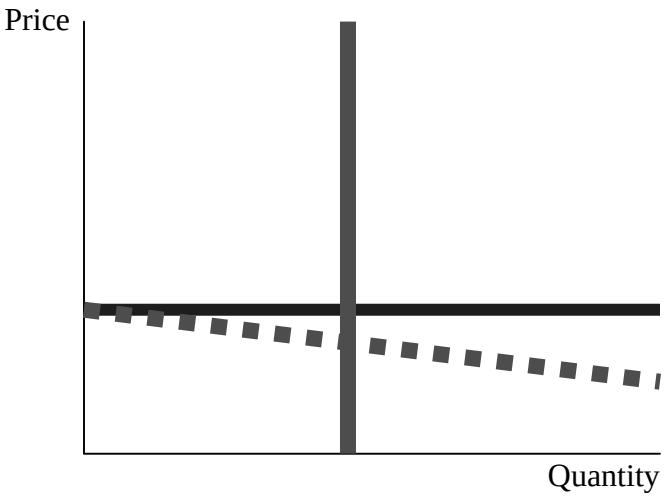


Figure 20.2 Supply and demand curve for stocks.

not affected by behavioral biases and that there are groups of investors who may sell stocks based upon faulty heuristics or mistaken beliefs. Early tests of whether demand curves for stocks are downwardly sloped focused on events in which the supply of the stock was independent of its economic value. For example, Andrei Shleifer (1986) examined all the cases in which index funds had to buy or sell a stock because it was added or deleted from the S&P 500 index. Since index funds seek only to mimic the index, rather than speculate on economic value, this provided a clean test of the sloped demand curve hypothesis. He found that the prices of stocks dropped from the index fell significantly when the change was made, while the prices of stocks added rose significantly—even though, in both cases, the fundamental value of the securities remained unchanged by their listing on the S&P 500. This study was important because it established that arbitrageurs are not able to entirely counterbalance the influence of large trades not motivated by economic valuation. Later studies using different indexes have shown this effect to be pervasive.

Although it is interesting to document the movement of a handful of stocks as they are listed or delisted from an index, what is the larger lesson of these findings? Could the demand curve for the stock market as a whole slope down? Could capital supply shocks—unrelated to economic valuation of the assets—move the aggregate price of the market? This is an important question because, put another way, it asks whether stock market bubbles can exist. To explore this issue, researchers have turned again to the S&P 500 index as a test case. For example, Goetzmann and Massa (2003) used daily inflows and outflows from three S&P 500 index funds as a measure of market demand and supply shocks by individual investors. They found that the market moved up on days when investors were buying and down on days they were selling. The evidence strongly suggested that these shocks affected prices rather than vice versa—unanticipated fund flows were correlated with S&P 500 returns in the last hour of trading but were uncorrelated to returns in the morning. Edelen and Warner (2001) documented this effect for the universe of equity mutual funds as a whole. Warther (1995) earlier found strong positive correlation between monthly equity fund flows and market returns.

Although the causal connection between price changes and inflows and outflows to the equity markets is now fairly well established, none of the studies discussed link this to investor psychology or mood. A few recent studies find a closer connection between fund flows and measures of investor beliefs. Indro (2004) finds that mutual fund flows are higher in the week following an increase in sentiment indicators collected by the American Association of Individual Investors and Investors Intelligence. Brown, Goetzmann, Hiraki, Otsuki, and Shiraishi (2001) use daily Japanese flows into "bull" and "bear" mutual funds and find that these sentiment measures are highly correlated to aggregate stock returns.

A number of studies have recently tried to determine if measures of market sentiment command a premium in expected return, as theorized by the costly (or risky) arbitrage model. For example, Brown et al., cited above, find strong evidence that the higher the exposure to the sentiment measure, the higher the realized return to an asset portfolio, controlling for traditional factors. Qiu and Welch (2004) find some evidence that the UBS/Gallup investor sentiment index explains cross-sectional differences in stock returns in the United States. Closed-end fund discounts were proposed by Lee, Shleifer, and Thaler (1991) as a sentiment indicator in an early attempt to test the limits to arbitrage model. Subsequent analysis of this variable, however, suggests it has little explanatory power in pricing models. Taking an agnostic approach about which variables capture sentiment, Baker and Wurgler (2006) try several different measures and find some cross-sectional explanatory power. They find that *positive sentiment is proxied* by six measures:[4]

1. small discounts (high premiums) on closed-end mutual funds; a discount is how much lower the net asset value of the funds is than their market value
2. high volume on the New York Stock Exchange
3. high first-day returns on initial public offerings (IPOs)
4. a high number of IPOs
5. high equity issuance relative to debt issuance
6. high market to book value for nondividend payers compared to payers

They also note that the stocks with exposure to sentiment variables are less liquid and harder to hedge—evidence that the arbitrageurs stay away from them.

The Marginal Investor

If arbitrageurs are not always the marginal investor in the market, then who sets prices? Researchers have found that in certain circumstances the characteristics of the marginal investor can change. The idea that the marginal investor in the market for an asset might belong to a specific clientele is not new. It was first documented empirically in the ex dividend behavior of stock prices. Elton and Gruber (1970, 2005), for example, found evidence that price changes around the dividend date for stocks were different depending on the size of the dividend and thus the tax characteristics of the stock. Clienteles, or subsets of participants in the capital markets, may also be defined by their behavior. Gompers and Metrick (1998) find that institutions are typically the marginal investors in the U.S. equity market. Griffin, Harris, and Topalogu (2006) look at trading during the NASDAQ bubble of the late 1990s and find evidence that momentum trades by sophisticated investors were correlated to market moves and thus identified as marginal. Goetzmann and Massa (2002) look at

[4]Barer and Wurgler construct a combined index from the measures using the first index from a principal component analysis.

index fund investors in a period in the 1990s and find that there are times when momentum traders appear to be the marginal investor, while other times contrarian investor actions are positively correlated to market-wide moves. These and other studies seem to indicate that the marginal investor may change through time and may reflect the motives of a specific subgroup of the investor populace—particularly in cases in which arbitrage is costly, risky, or difficult. The most important case in which arbitrage is particularly risky is when the market as a whole deviates from its absolute value. Choosing to bet against the broad market trend means taking an unhedged, contrarian position and waiting for the rest of the investment world to come around to your view. It would have been difficult to hold a short position in the NASDAQ from 1997 to 2000, regardless of the strength of one's conviction. Thus deviation of the market as a whole from fundamental value is easier to justify logically than deviations of prices of individual, hedgeable securities from the prices of close economic substitutes.

Stock Prices and Social Dynamics

This idea of a psychologically induced market-wide disjunction in economic value has a long history. MacKay's (1841) classic *Extraordinary Popular Delusions and the Madness of Crowds* attributed both the Dutch tulip bubble of the seventeenth century and the South Seas Bubble of 1720 to feverish and foolish investor behavior. More recently, Robert Shiller's book *Irrational Exuberance* forecast the bursting of the millennium tech bubble by attributing much of the increase in technology stocks to popular irrational optimism about their long-term earnings potential.

These age-old themes have provided the fundamental motivation for research on the issue of whether investor sentiment is able to move the market as a whole. This was formally raised by Robert Shiller (1981), who argued that social and psychological factors had the potential to affect stock prices in a significant way—at times driving them far away from rational, economic values. His theory, developed over several research studies, is one of the most widely debated studies in behavioral finance. Shiller's basic claim is that the stock prices move around more than can be justified by changing expectations of future dividend flows. In an elegant and simple argument, he noted that stock prices are expectations of discounted future dividends. Because an expectation of a variable must have a lower volatility than the variable itself, stock prices should fluctuate less than a series of discounted future dividends. Instead, his analyses indicated that stocks were significantly more volatile than discounted future streams of historically realized dividends. In the 25 years since the excess volatility hypothesis was first proposed, a number of studies have pointed out problems in the test methodology and in the interpretation of the results, while others have provided supporting evidence. The initial proposition that investor psychology could potentially explain a significant component of asset returns has stimulated considerable future inquiry and debate; however, more direct tests of the influence of investor thoughts and beliefs were impossible without the collection of behavioral data.

Recognizing the need for direct behavioral data to test his theory, Shiller began to poll investors on a regular basis in the United States and Japan regarding their expectations about future stock returns shortly after the crash of 1987. A chart from the Shiller Investor Confidence Survey is shown in Figure 20.3; the lines indicate the market sentiment of both institutional and individual investors. Investor outlook for the stock market indeed fluctuated significantly through time, consistent with his theory. Institutional investors generally agreed with the assessment of individual investors, although individuals were noticeably more bullish in 2001 and 2002—perhaps hopeful the market might still rebound from its crash around

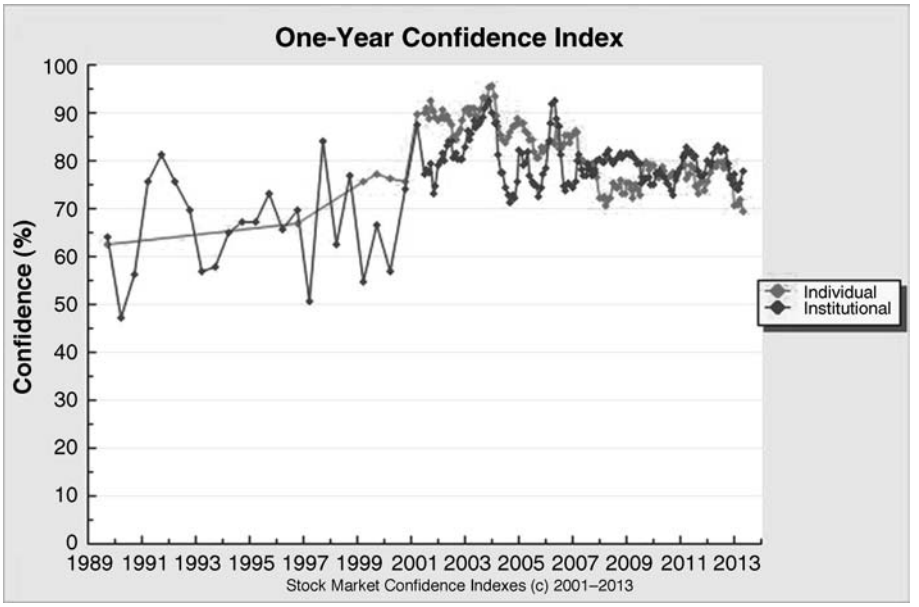


Figure 20.3 Time series of one year confidence intervals. *Source*: International Center for Finance (http://icf.som.yale.edu/stock-market-confidence-indices-united-states-yearindex).

the turn of the millennium. Did a sentiment shift among individual investors cause stock prices to plummet in the year 2000? Figure 20.3 suggests otherwise. Institutional investor sentiment appears to have been at a local low point in 1999 but rose in the year 2000—it is tempting to attribute this to beliefs adjusting to past decisions.

In sum, much of the empirical research linking asset prices to investor behavior over the past few decades has focused on the question of whether security prices can deviate from fundamental economic values as a result of actions by unsophisticated investors as opposed to well-informed, astute speculators. The limits to arbitrage theory show how this can happen, and considerable empirical evidence linking trades by retail investors to stock price dynamics at the aggregate level suggests that it does happen. Studies of behavioral "factors" constructed from flows and other sentiment indicators appear to show that they can also lead to differences in realized returns among different classes of equities. This latter result is suggestive—but not conclusive—evidence that investors require a premium for stocks exposed to a sentiment factor. The age-old debate about whether stock market bubbles and crashes are driven by investor sentiment is still not settled, despite serious attempts to collect sentiment data that would allow such a test. At the current time, the theory cannot be ruled out, but neither can it be rigorously tested.

Media and Behavior and Contrarian Investors

Current research on investor sentiment and market prices has delved deeply into the role played by the news media. For sentiment to have an effect on prices, there must be some coordination mechanism that focuses investor attention on a particular stock and causes reading that moves the stock price. To test this, Tetlock (2007) collected news items from the *Wall Street Journal* about publicly traded companies and coded them according to their fraction of positive or negative words to analyze the implicit sentiment. He found that high

sentiment—positive and negative—predicted volume of trade.[5] Barber and Odean (2008) find that stocks in the news capture the attention of individual retail investors and cause them to buy. Dougal, Engelberg, Garcia, and Parsons (2012) find that the sentiment expressed by journalists can actually move broad indexes. In a clever study of the rotation of contributors to the *Wall Street Journal*'s "Abreast of the Market" column, they found that optimistic articles preceded short-term gains in the Dow and that pessimistic articles preceded short-term losses. The importance of sentiment in the media also extends to the Internet. Da, Engelberg, and Gao show that Google Internet searches for stocks predict stock price increases over the following two weeks, followed by price reversal. This is consistent with temporary demand pressure due to investor attention. Short media bursts lead to temporary mispricings, indicative of sentiment-driven price variation and establishing the importance of behavioral finance to the field of asset pricing. While investor attention and trading seems responsive to media, it does not always lead to losses. Kaniel, Saar, and Titman (2008) discovered that small investors appeared to be contrarian—buying after price declines and selling after price increases. When these actions were particularly intense, investors profited from price reversals. This contrarian behavior is consistent with other research using microstructure data and investor questionnaires. Kelley and Tetlock (2012) also find evidence that passive buyers after price declines profit, perhaps because they are liquidity providers.

Explaining Anomalies

One direct outcome of the behavioral finance paradigm has been scholarly excitement about its potential to explain previously documented stock market anomalies. Over many years, empirical researchers in finance had documented a number of apparent violations of efficient market theory. Usually these violations take the form of risk-adjusted returns generated by back-tests of trading rules applied to U.S. stock market data. These apparent violations include market seasonality, market underreaction to news such as earnings announcements, size-related return differentials, and return differentials associated with financial ratios such as price/book value and price/earnings.

Financial economists have come up with a variety of theoretical models to show how these empirical anomalies could be due to the effects of loss aversion, prospect theory, overconfidence, and other kinds of psychological heuristics. These heuristics were identified by Daniel Kahneman and Amos Tversky. They found, for example, that investors were more adverse to recognizing a loss of a certain quantity than recognizing a gain (*loss averse*), investors tended to compare their gains and losses to a benchmark—such as what they originally paid for an item (*Prospect Theory)*, and investors appeared more confident of their estimates of uncertain values than was warranted by the statistical characteristics of the data (*overconfident*). These tendencies led to poor or contradictory choices by the experimental subjects. Financial researchers have asked whether such tendencies could explain known anomalies or irregularities in market behavior. Barberis, Huang, and Santos (2001), for example, asked how the existence of investors with utility functions consistent with Prospect Theory would affect asset prices. Daniel, Hirshleifer, and Subrahmanyam (2001) explore the effect of investor overconfidence on the covariates of asset returns. Many other researchers have developed models to explain pricing anomalies with Kahneman and Tversky–like heuristics.

In an early empirical study of one such model, Werner De Bondt and Richard Thaler (1985) tested the implications of the representativeness heuristic in what has become a

[5]Tetlock (2007); Dougal, Engelberg, Garcia, and Parsons (2012)

widely influential study of stock price overreaction. The "representativeness" heuristic is the idea that investors will overreact to recent information, treating recent news as more relevant than it actually is for forecasting future performance. Using historical information on the returns to individual securities in the U.S. market, the authors formed portfolios of securities that had recently decreased in value and offset this investment with a short position in stocks that had recently increased in value. They found that this back-tested strategy yielded consistent positive risk-adjusted returns—precisely the pattern one would expect if investors prone to a representativeness heuristic were influencing prices. While the mean reversion documented by De Bondt and Thaler and several other scholars is undeniably present in the return data, without actual investor trading data, it was difficult at the time of the study to draw a direct causal link between investor response to information and stock price changes.

The disposition effect is another heuristic that has motivated empirical tests. Grinblatt and Han (2004) show that the disposition effect (the tendency to ride losers and sell winners) explains much of the well-known momentum profits to riding winners at the 12-month horizon. They rely on past volume and price changes to identify stocks with a "disposition" hangover. Goetzmann and Massa (2002) use actual purchases and sales by individual investors to identify stocks with a disposition "hangover" and find results consistent with Grinblatt and Han. Although institutional investors are typically regarded as exempt from the disposition effect, Frazzini (2004) shows how the trades of disposition-prone mutual funds can be used to generate profits, at least in back-tests.

Two problems currently confront the attempts to explain asset pricing anomalies with behavioral models. The first, a widely recognized problem, is that there is no single, consistent model of investor behavior proposed by researchers in behavioral finance that may be falsified. Thus, while classical theories such as the capital asset pricing model have unambiguous empirical predictions, most behavioral models do not. Investor overreaction is consistent with one type of investor heuristic, while overconfidence is consistent with another. This problem can be interpreted as a sign that the field of behavioral finance, despite 25 years of exciting research, has not yet developed a complete, internally consistent, testable model of investor cognition and action.

The second problem is that the majority of empirical studies in the area of behavioral finance do not use behavioral data. The use of stock price data to prove that investor psychology affects stock prices is nearly tautological. A test of whether investor psychology may influence behavior and whether this behavior in turn may influence prices requires different and richer data and a considerable burden of proof. To this end, it is incumbent upon scholars in the field of behavioral finance to collect and use behavioral data.

BIBLIOGRAPHY

1. Akerlof, George A., and Dickens, William T. "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, **72,** No. 3 (1982), pp. 307–319.
2. Baker, Malcolm, and Wurgler, Jeffrey. "Investor Sentiment and the Cross-Section of Stock Returns," *Journal of Finance*, **61,** No. 4 (2006), pp. 1540–1626.
3. Barber, Brad M., and Odean, Terrance. "All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors," *Review of Financial Studies*, **21** (2008), pp. 785–818.
4. Barberis, Nicholas, Huang, Ming, and Santos, Tano. "Prospect Theory and Asset Prices," *Quarterly Journal of Economics*, **XVI,** No. 1 (2001), pp. 1–53.
5. Benartzi, Shlomo, and Thaler, Richard. "How Much Is Investor Autonomy Worth?" UCLA Anderson School working paper (2001).

6. Beshears, John, Choi, James J., Laibson, David, and Madrian, Brigitte C. "The Importance of Default Option for Retirement Saving Outcomes: Evidence from the United States," in *Social Security Policy in a Changing Environment* (Chicago: University of ChicagoPress, 2009).
7. Bracha, Anat. "Affective Decision Making in Insurance Markets," International Center for Finance, Yale School of Management working paper No. 04-03 (2004).
8. ——. "Consistency and Refutability of Affective Choice," International Center for Finance, Yale School of Management working paper No. 04-40 (2004).
9. Bris, Arturo, Goetzmann, William N., and Zhu, Ning. "Efficiency and the Bear: Short Sales and Markets around the World," *Journal of Finance*, **LXII** (2007), pp. 1029–1079.
10. Brown, Stephen, Goetzmann, William, Hiraki, Takato, Otsuki, Toshiyuki, and Shiraishi, Noriyoshi. "The Japanese Open-End Fund Puzzle," *Journal of Business*, **74** (2001), pp. 59–77.
11. Chen, Keith, Santos, Laurie, and Lakshminarayanan, Vankat. "The Evolution of Our Preferences: Evidence from Capuchin-Monkey Trading Behavior," Yale School of Management working paper (2005).
12. Choi, James J., Laibson, David, Madrian, Brigitte C., and Metrick, Andrew. "Defined Contribution Pensions: Plan Rules, Participant Choices, and the Path of Least Resistance," *Tax Policy and the Economy*, **XVI** (2002), pp. 67–113.
13. Choi, James J., Coval, Joshua D., and Moskowitz, Tobias J. "Home Bias at Home: Local Equity Preference in Domestic Portfolios," *Journal of Finance*, **54,** No. 6 (Dec. 1999), pp. 2045–2073.
14. ——. "The Geography of Investment: Informed Trading and Asset Prices," *Journal of Political Economy*, **109,** No. 3 (2001), pp. 811–841.
15. Daniel, Kent D., Hirshleifer, David, and Subrahmanyam, Avanidhar. "Overconfidence, Arbitrage, and Equilibrium Asset Pricing," *Journal of Finance*, **56,** No. 3 (June 2001), pp. 73–84.
16. De Bondt, Werner F. M., and Thaler, Richard. "Does the Stock Market Overreact?," *Journal of Finance*, **40,** No. 3 (Jul. 1985), pp. 793–805.
17. Dhar, Ravi, and Zhu, Ning. "Up Close and Personal: An Individual Level Analysis of the Disposition Effect," *Management Science*, **52,** No. 3 (2006), pp. 726–740.
18. Dichev, Ilia D., and Janmes, Troy. "Lunar Cycle Effects in Stock Returns," *Journal of Private Equity*, **6,** No. 3 (Fall 2003), pp. 8–29.
19. Dougal, Casey, Engelberg, Joseph, Garcia, Diego, Parsons, Christopher A. "Journalists and the Stock Market," *Review of Financial Studies*, **25,** No. 3 (2012), pp. 639–679.
20. Doukas, John A., and Milonas, Nikolaos T. "Investor Sentiment and the Closed-End Fund Puzzle: Out-of-Sample Evidence," *European Financial Mangement*, **10,** No. 2 (Dec. 2002), pp. 220–230.
21. Edelen, R., and Warner, J. "Aggregate Price Effects of Institutional Trading: A Study of Mutual Fund Flow and Market Returns," *Journal of Financial Economics*, **59** (Feb. 2001), pp. 195–220.
22. Elton, Edwin, and Gruber, Martin. "Marginal Stockholder Tax Rates and the Clientel Effect," *Review of Economics and Statistics*, **52,** No. 1 (1970), pp. 68–74.
23. Elton, Edwin J., and Gruber, Martin J. "Marginal Stockholder Tax Effects and Ex Dividend Day Price Behavior: Evidence from Taxable versus Non-taxable Closed-end Funds," *Review of Economics and Statistics*, **87,** No. 3 (Aug. 2005), pp. 579–586.
24. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher. "Participant Reaction and the Performance of Funds Offered by 401K Plans," Stern School of Business, NYU working paper (2006).
25. Elton, Edwin J., Gruber, Martin J., and Rentzler, Joel C. "Professionally Managed, Publicly Traded Commodity Funds," *Journal of Business*, **60,** No. 2 (April 1987), pp. 220–231.
26. Festinger, Leon. *A Theory of Cognitive Dissonance* (Stanford, CA: Stanford University Press, 1957).
27. Frazzini, Andrea. "The Disposition Effect and Under-reaction to News," International Center for Finance, Yale School of Management working paper No. 04–24 (July 2004).
28. Goetzmann, W. N., and Mark, G. "Does Delisting from the S&P 500 Affect Stock Price?" *Financial Analysts Journal*, **42** (1986), pp. 64–69.

29. Goetzmann, W. N., and Massa, Massimo. "Daily Momentum and Contrarian Behavior of Index Fund Investors," *Journal of Financial and Quantitative Analysis*, **37** (Sept. 2002), pp. 375–389
30. ——. "Index Funds and Stock Market Growth," *Journal of Business*, **76,** No. 1 (2003), pp. 1–27.
31. Goetzmann, W. N., and Peles, N. "Cognitive Dissonance and Mutual Fund Investors," *Journal of Financial Research*, **20,** No. 2 (Summer 1997), pp. 145–158.
32. Goetzmann, W. N., and Zhu, Ning. "Rain or Shine: Where Is the Weather Effect?" *European Financial Management*, **11** (2005), pp. 559–578.
33. Goetzmann, William N., and Dhar, Ravi. "Bubble Investors: What Were They Thinking?" working paper, Yale University (2006).
34. Gompers, Paul A., and Metrick, Andrew. "Institutional Investors and Equity Prices," NBER working paper No. W6723 (Sept. 1998).
35. Griffin, John, Harris, Jeffrey, and Topalogin, Selim. "Who Drove and Burst the Tech Bubble," University of Texas working paper (2006).
36. Grinblatt, Mark, and Han, Bing. "Prospect Theory, Mental Accounting, and Momentum," International Center for Finance, Yale School of Management working paper No. 00–7 (Aug. 2004).
37. Grinblatt, Mark, and Keloharju, Matti. "What Makes Investor Trade?" *Journal of Finance*, **56,** No. 2 (2001a), pp. 589–616.
38. ——. "How Distance, Language, and Culture Influence Stock Holdings and Trades," *Journal of Finance*, **56** (2001b), pp. 1053–1073.
39. Hirshleifer, David, and Shumway, Tyler. "Good Day Sunshine: Stock Returns and the Weather," *Journal of Finance*, **58** (2003), pp. 1009–1032.
40. Huberman, Gur. "Familiarity Breeds Investment," *Review of Financial Studies*, **14** (2001), pp. 659–680.
41. Indro, Daniel. "Does Mutual Fund Flow Reflect Investor Sentiment?" *Journal of Behavioral Finance*, **V,** No. 2 (Aug. 2004), pp. 105–115.
42. Ivkovic, Zoran, and Weisbrenner, Scott J. "Local Does as Local Is: Information Content of the Geography of Individual Investors' Common Stock Investments," University of Illinois at Urbana-Champaign working paper (2003).
43. Jones, Charles M., and Lamont, Owen A. "Short Sale Constraints and Stock Returns," *Journal of Financial Economics*, **66,** No. 2 (2002), pp. 207–239
44. Kahneman, Daniel, and Tversky, Amos. "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, **47,** No. 2 (March 1979), pp. 263–292.
45. Kamstra, Mark J., Kramer, Lisa A., and Levi, Maurice D. "Winter Blues: A SAD Stock Market Cycle," Federal Reserve Bank of Atlanta working paper No. 2002–13a, http://ssrn.com/abstract=208622 (Oct. 2003).
46. Kaniel, Ron, Saar, Gideon, and Titman, Sheridan. "Individual Investor Trading and Stock Returns," *Journal of Finance*, **63,** No. 1 (2008), pp. 273–310.
47. Keley, Eric, and Tetlock, Paul. "How Wise Are Crowds? Insights from Retail Orders and Stock Returns," *Journal of Finance*, **68,** No. 3, pp. 1229–1265.
48. Kumar, Alok. "Is the Local Bias of Individual Investors Induced by Familiarity or Information-Asymmetry?" Notre Dame University working paper (2005).
49. Lamont, Owen, and Thaler, Richard. "Can the Market Add and Subtract? Mispricing in Tech Stock Carve-outs," *Journal of Political Economy*, **111,** No. 2 (April 2003), pp. 280–292.
50. Lee, Charles, Shleifer, Andrei, and Thaler, Richard. "Investor Sentiment and the Closed-End Fund Puzzle," *Journal of Finance*, **46,** No. 1 (1991), pp. 75–109.
51. Linnainmaa, Juhani T. "Do Limit Orders Alter Inferences about Investor Performance and Behavior?" *Journal of Finance*, **65,** No. 4 (2010), pp. 1473–1506.
52. Lo, Andrew, and Repin, Dmitry V. "The Psychophysiology of Real-Time Financial Risk Processing," *Journal of Cognitive Neuroscience*, **14,** No. 3 (2002), pp. 323–339.
53. Loughran, Tim, and Schultz, Paul. "Weather, Stock Returns, and the Impact of Localized Trading Behavior," *Journal of Financial and Quantitative Analysis*, **39** (June 2004), pp. 343–364.
54. Loewenstein, George, and O'Donoghue, Ted. "Animal Spirits: Affective and Deliberative Processes in Economic Behavior," CAE working paper 04-14 (2004).

55. Mackay, Charles. "Extraordinary Popular Delusions and the Madness of Crowds." (1841), reprinted by Barnes and Noble (2004).
56. Markowitz, Harry. "The Utility of Wealth," *Journal of Political Economy*, **60,** No. 2 (April 1952), pp. 151–158.
57. Massa, Massimo, and Simonov, Andrei. "Behavioral Biases and Investment," *Review of Finance*, **9,** No. 4 (2005), pp. 483–507.
58 Odean, Terrance. "Are Investors Reluctant to Realize Their Losses?" *Journal of Finance*, **53,** No. 5. (Oct. 1998), pp. 1775–1798.
59. Qui, Lily, and Welch, Ivo. "Investor Sentiment Measures," NBER working paper No. W10794 (Sept. 2004).
60. Rashes, Michael S. "Massively Confused Investors Making Conspicuously Ignorant Choices (MCI-MCIC)," *Journal of Finance*, **56,** No. 5 (2001), pp. 1911–1927.
61. Ross, Stephen A. "The Arbitrage Pricing Theory of Capital Asset Pricing," *Journal of Economic Theory*, **13,** No. 3, (1976), pp. 341–360.
62. Saunders, Edward M. "Stock Prices and Wall Street Weather," *American Economic Review*, **83,** (1993), pp. 1337–1345.
63. Shiller, Robert J. "Do Stock Prices Move Too Much to Be Justified by Subsequent Changes in Dividends?" *American Economic Review*, **71** (June 1981), pp. 421–436.
64. ——. "Measuring Bubble Expectations and Investor Confidence," *Journal of Psychology and Financial Markets*, **1,** No. 1 (2000), pp. 49–60.
65. _____. *Irrational Exuberance*. Princeton University Press: Princeton New Jersey (2000).
66. Shefrin, Hersh, and Statman, Meir. "The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence," *Journal of Finance*, **40,** No. 3 (July 1985), pp. 777–790.
67. ——. "Behavioral Portfolio Theory," *Journal of Financial and Quantitative Analysis*, **35,** No. 2 (June 2000), pp. 480–502.
68. Shiv, Baba, Loewenstein, George, Bechara, Antoine, Damasio, Hanna, and Damasio, Antonio R. "Research Report, Investment Behavior and the Negative Side of Emotion," *Psychological Science*, **16,** No. 6 (June 2005), pp. 435–439.
69. Shleifer, Andrei. "Do Demand Curves for Stocks Slope Down?" *Journal of Finance*, **41,** (1986), pp. 579–590.
70. Shleifer, Andrei, and Vishny, Robert W. "The Limits of Arbitrage," *Journal of Finance*, **52,** No. 1 (March 1997), pp. 35–55.
71. Tversky, Amos, and Kahneman, Daniel. "Judgment under Uncertainty: Heuristics and Biases," *Science*, New Series, **185,** No. 4157 (Sep. 1974), pp. 1124–1131.
72. ——. "The Framing of Decisions and the Psychology of Choice," *Science*, New Series, **211,** No. 4481 (Jan. 1981), pp. 453–458.
73. Yuan, Kathy, Zheng, Lu, and Zhu, Qiaoqiao, "Are Investors Moonstruck? Lunar Phases and Stock Returns," *Journal of Empirical Finance*, **13,** No. 1 (2006), pp. 1–23.
74. Warther, Vincent. "Aggregate Mutual Fund Flows and Security Returns," *Journal of Financial Economics*, **39,** No. 2–3 (1995), pp. 209–235.
75. Zhu, Ning. "The Local Bias of Individual Investors," International Center for Finance, Yale School of Management working paper (2005).

21

Interest Rate Theory and the Pricing of Bonds

Until the last few decades, bond valuation was considered a rather dull subject. After all, a bond is easier to value than a stock because the issuer has agreed to a certain stream of payments (coupon and principal) and the bond has a maximum life (maturity).

Two factors led to a change in the difficulty of valuation. First, the timing of cash flows became more variable and their payment less certain because new types of instruments were issued. For example, bonds were issued with more complex options, which could affect both the timing and magnitude of the cash flows. In addition, more risky debt was issued with less certain cash flows. Second, valuation became more difficult because interest rates become more volatile. When interest rates go up, bond prices fall so that outstanding bonds offer returns similar to those earned by new issues. Interest rates were volatile during the 1970s and the 1980s. Accompanying this increased volatility were huge swings in the market value of bond portfolios. This increased volatility in market values was viewed as an opportunity and as a risk. Active bond portfolio management began to receive a lot of attention.

Table 21.1 presents the yearly holding period return that would have been earned by holding three different portfolios of bonds from 1999 to 2011. Returns from holding long-term government bonds were extremely volatile during this period. For example, the return from this portfolio was −14.990 in 2009 and 25.990 in 2008. These returns bear little resemblance to the interest rate on long-term bonds during those years. Given the variability of bond returns, you might suspect that we are heading toward a consideration of a portfolio theory for bonds. In fact, that is the subject of the next chapter. But before we attempt to construct portfolio strategies, we must understand the pricing of bonds, which is discussed in this chapter. The first part of this chapter, after briefly introducing the major types of bonds, discusses the many meanings of interest rates and

Table 21.1 Rates of Return on Selected Bond Portfolios

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Int gov	−1.8	12.6	7.6	12.9	2.4	2.3	1.4	3.1	10.1	13.1	−2.4	7.1	9.5
Long gov	−9.0	21.5	3.7	17.8	1.4	8.5	7.8	1.2	9.9	25.9	−14.9	10.1	28.2
Long corp	−7.4	12.9	10.6	16.3	5.9	8.7	5.9	3.2	2.6	8.8	3.0	12.4	17.9

places special emphasis on the role of one of these rates, the spot rate, in determining bond prices. The second part of the chapter discusses the determination of bond prices. The third and longer part of the chapter deals with the factors that explain bond prices.

AN INTRODUCTION TO DEBT SECURITIES

Bonds are primarily traded over the counter rather than on organized exchanges. For some bond issues, the bond markets are highly liquid. Other bonds rarely trade. There are four major categories of long-term fixed income securities:

1. federal government bonds
2. corporate bonds
3. mortgages
4. municipal bonds

Although these were discussed in Chapter 2, we review their major characteristics here.

Government Bonds

Government bonds represent the borrowing of the federal government. They represent the largest percentage of the total debt market and are by far the most liquid. Because they are backed by the government, they are considered default free. They are the simplest to value; they pay interest at a fixed rate and have a stated principal.

Corporate Bonds

Corporate bonds are debt obligations of corporations. Corporate issues can be publicly traded or privately placed, usually with a bank or insurance company. The publicly traded corporate market is much less active than the government market, with many of the issues rarely if ever trading after the initial offering. Corporate bonds are backed by the credit of the issuing corporation. It is the corporation's ability to earn money and meet the obligations of the debt issue that determines the bond's default risk. Generally, corporate bonds are divided into investment grade, where the risk of default is low, and high yield or junk, where the risk of default is substantial. Although many different types of option features can be present on corporate bonds, callability, sinking funds, and convertibility are the most common. A call provision on a bond gives the issuing corporation the right to force the bondholder to sell the bond back to the corporation at a particular price. The price is known and may vary over time. The right to call rests with the corporation, and hence callable bonds must offer a higher return to compensate the holder for a disadvantageous call. Sinking funds options are like the call option. A large corporate bond issue may have a sinking fund provision. Consider a 100 million 10-year bond issue. The sinking fund provision may require the corporation to retire 10 million in face value of bonds a year for each of the 10 years. The sinking fund is intended to prevent the corporation from having to make one large repayment. The corporation generally has the option of purchasing the bonds in the open market or calling them back from the investors. Thus the investor risks having the bond called back to meet the sinking fund. Once again, because the corporation has the option, the investor will require a higher return, everything else held constant, to compensate for the disadvantageous call; everything else is not constant, however. The presence of a sinking fund lowers the risk that the firm will default on the entire issue of bonds or perhaps on any of it. The presence of a sinking fund

with the ability to call bonds to meet it results in a lower default risk but a higher interest rate risk for the bondholder. Finally, convertible bonds are bonds that can be exchanged for another security, usually common equity. Because the option rests with the bondholder, the bondholder has a potentially valuable option, and these bonds are issued with lower interest payments than nonconvertible bonds.

Mortgage Bonds

Mortgages are debt obligations backed by real estate. Most mortgages are originated by a bank, insurance company, or other financial institution. Then many mortgage loans are publicly traded by pooling a group of mortgages and issuing bonds against the pool. The most liquid of the publicly traded mortgage instruments are Ginnie Maes, which are bonds backed by a pool of mortgages. The Ginnie Mae insures the payment of principal and interest on the mortgages and extracts a fee for the insurance. In addition, a fee for collection of the mortgage payments is extracted. The remainder of the principal and interest payments on the mortgage is passed along to the Ginnie Mae owner. Because mortgages are paid monthly, interest on Ginnie Maes is also paid monthly. Ginnie Maes have interesting risk characteristics in that the default risk has been removed by the issuer, but there is major risk from the uncertainty associated with the timing of the payment stream. Homeowners have the option to prepay their mortgages, and they will generally do so if they sell the home or if interest rates fall sufficiently. Because these options rest with the payer of the mortgage, the purchaser of a Ginnie Mae is uncertain about the size of the payment that will be received. To compensate for this uncertainty, investors in Ginnie Maes will require a higher return than on comparable governments.

Municipal Bonds

The final major category of bonds is municipals. These are debt obligations of states, cities, and state or city authorities. Municipal bonds are generally divided into two broad categories—bonds that are backed by the full faith and credit of the city or state and those that are backed by a government agency or authority. The latter, called revenue bonds, would be issued by a government agency such as a port authority or turnpike authority and are backed by the revenues generated by the agency. Municipal bonds have default risk. Their major distinguishing characteristic is that the interest on municipal bonds is exempt from federal tax and sometimes from state tax, depending on the issuing state and the residence of the purchaser.

THE MANY DEFINITIONS OF RATES

An investor who examines the literature on bond valuation will find a confusing array of terms all seemingly related to interest rates—terms like *spot rates*, *future rates*, *yield to maturity*, and *current yield*. In the following, we define and explain these alternative rates.

The rate most investment professionals use to compare bonds is yield to maturity. The method used to calculate the yield to maturity varies across bond categories. Thus yield to maturities on different types of instruments may not be comparable. In what follows, we discuss general principles underlying the calculation of yield to maturity, the variations in calculations across bond categories, and how to make the calculations comparable. The yield to maturity is the internal rate of return earned from holding a bond to maturity. The yield to maturity on a three-year bond with annual interest payments of $100, a principal payment of $1,000, and a cost of $900

is that rate (y) that equates the present value of the three cash flows on the bond with its present price or[1]

$$900 = \frac{100}{(1+y)} + \frac{100}{(1+y)^2} + \frac{100+1000}{(1+y)^3}$$

Therefore

$$y = 14.3\%$$

This expression for yield to maturity can also be written in summation notation. Let $C(t)$ be the cash flow in t. The cash flow in the example is either the coupon of \$100 or the principal plus interest of \$1,100. Then, in summation notation, yield to maturity is the value of y that solves the following expression:

$$\text{Price} = \sum_t \frac{C(t)}{(1+y)^t}$$

The frequency of compounding assumed in computing the yield to maturity varies across types of bonds. We review several compounding conventions here.

Government bonds and notes and most corporate bonds pay interest semiannually. The yield to maturity on these bonds is calculated differently from the earlier example. Assume a three-year bond with semiannual interest payments of \$50, a principal payment of \$1,000, and a cost of \$900. The yield to maturity is calculated as follows:

$$900 = \frac{50}{\left(1+\frac{y}{2}\right)} + \frac{50}{\left(1+\frac{y}{2}\right)^2} + \frac{50}{\left(1+\frac{y}{2}\right)^3} + \frac{50}{\left(1+\frac{y}{2}\right)^4} + \frac{50}{\left(1+\frac{y}{2}\right)^5} + \frac{1050}{\left(1+\frac{y}{2}\right)^6}$$

Thus

$$y = 14.2\%$$

The yield to maturity calculated in this way is also called the *bond equivalent yield*. This method of determining the yield to maturity is based on a rather arbitrary assumption about reinvestment. Although it assumes discounting and compounding on a semiannual basis, it assumes no compounding in converting semiannual yield to an annual yield. That is, the semiannual rate of return is converted to an annual return by multiplying it by 2. This ignores the fact that the investor can earn interest on the first coupon received any year for the second half of the year. If one assumes that interest can be earned on the first payment received in a year, then the actual annual return is the value at the half year $(1 + y/2)$ times the return in the second half year $(1 + y/2)$ or on the example $[(1.071)^2] - 1 = 14.7\%$. This is often called the effective annual yield (y_E), and it is always higher than the yield to maturity stated on the bond.

The effective annual yield represents the annual return the investor will receive if she holds the bond to maturity and if coupons are reinvested every six months at one-half the bond equivalent yield for each six-month period.[2] Similar methodology and terminology apply for debt instruments that pay interest at more frequent intervals than semiannually. For example,

[1]Eurobonds, which are bonds not registered with the Securities and Exchange Commission (SEC), pay annual interest, and their yield to maturity is calculated in the manner shown subsequently.

[2]The reader should be alerted to the fact that the quoted price on bonds is not the trade price. The trade price includes accrued interest. (See Appendix A.)

Ginnie Maes have monthly cash flows of interest and principal. A 30-year Ginnie Mae would have 360 payments (12 × 30). If $C(t)$ is the payment, then the yield to maturity is y, where

$$\text{Price} = \sum_{t}^{360} \frac{C(t)}{\left(1+\frac{y}{12}\right)^{t}}$$

In computing the yield to maturity, the monthly interest rate $y/12$ is annualized by multiplying by 12. Once again, no compounding is assumed in annualizing. The effective annual yield is one plus the monthly interest rate to the 12th power:

$$y_E = \left(1+\frac{y}{12}\right)^{12} - 1$$

For example, if $C(t)$ is \$8,482 and the price is \$1,000,000, then

$$1{,}000{,}000 = \sum_{t}^{360} \frac{8482}{\left(1+\frac{y}{12}\right)^{t}}$$

and the yield to maturity is

$$y = 9.6\%$$

The effective annual yield is

$$y_E = \left(1+\frac{0.096}{12}\right)^{12} - 1$$

$$y_E = 10.03\%$$

The quoted yield on Treasury bills is computed very differently than quoted yields on other instruments. Because Treasury bills are an important instrument, it is worthwhile discussing how rates are calculated.

Treasury bills are government debt issued with maturities of one year or less. There are only two cash flows associated with Treasury bills, one with the original purchase and one when the Treasury bill matures (they do not pay interest). A Treasury bill with a maturity of 60 days may be issued at 99 and mature at 100. The return is earned by the appreciation from 99 to 100. The interest rate on Treasury bills is calculated by the following formula, called the bankers' discount yield:

$$b = \frac{P_1 - P_0}{P_1} \cdot \frac{360}{N}$$

where

P_1 is ending price

P_0 is beginning price

N is number of days to maturity

In the preceding example, the bankers' discounted yield would be

$$b = \frac{100-99}{100} \times \frac{360}{60} = 6.00\%$$

As we have shown, the method used to calculate yield to maturity varies across instruments. Those investors using yield to maturity to compare bonds should adjust the calculations so that a common set of assumptions is being used. This is true when comparing Treasury bills to other government bonds. This is also true when comparing Ginnie Maes or Eurobonds to governments. Most institutions either calculate the effective annual yield on all instruments or adjust all instruments to have the same assumptions as government bonds by calculating a semiannual interest rate and doubling it (the bond equivalent yield).[3] Methods for doing this are presented in Appendix C.

Although the yield to maturity is the most common rate used in the investment community, there are problems with it. The yield to maturity is the return if all cash flows received before the horizon are invested at the yield to maturity to the horizon. Because different bonds have different yield to maturities, an investment organization choosing among bonds with different yields to maturity is making different assumptions concerning the reinvestment rate.

As an illustration of the difficulty this causes, consider the following example:

	Bond *A*	Bond *B*
Coupon	10%	3%
Principal	100	100
Price	$138.90	$70.22
Maturity	15 years	15 years
Frequency of payment	Annual	Annual
Yield to maturity	6%	6.1%

In calculating the yield to maturity, the implicit assumption is that cash flows are reinvested at 6% for bond *A* and 6.1% for bond *B* (the respective yield to maturities).

For an organization, there will be some rate at which funds are invested, and this will be the same rate no matter which bond the coupon payments come from. For any reinvestment rate above 6.43% the value in 15 years will be higher for bond *A* than for bond *B*.[4]

[3]For example, for Ginnie Maes, one would take the monthly interest rate $y/12$ and compute the semiannual interest rate

$$\left[\left(1+\frac{y}{12}\right)^6-1\right]$$

The semiannual interest rate is then doubled to get an annual rate. Similarly for Treasury bills, the return earned over the life of the Treasury bill can be calculated by

$$r = b \cdot N/360 \cdot \frac{P_1}{P_0}$$

and the bond equivalent yield is

$$2\times\left[(1+r)^{365/2N}-1\right]$$

See Appendix C for the calculations for all instruments.

[4]In the next section we will show that the price of a bond is determined by discounting the cash flows at spot rates. In calculating the prices, a sharply rising yield curve was assumed with subsequent one-period rates above 5.9% throughout and in fact above 7.7% by period 2. Thus the anticipated reinvestment rate is well above 6.43%, and the organization should prefer bond *A*.

Table 21.2 Illustrating the Nonadditivity of Yields

Outlay Bond	(Price)	Periods 1	2	3	Yield to Maturity	Weighted Average Yield
A	−100	15	15	115	15.00%	
B	−100	6	106		6.00%	
C	−92	9	9	109	12.35%	
A + *B*	−200	21	121	115	11.29%	10.50%
B + *C*	−192	15	115	109	9.65%	9.04%
A + *C*	−192	24	24	224	13.71%	13.73%

In addition, because of the differing reinvestment assumptions, yields are not additive. The yield to maturity on a portfolio is not a weighted average of the yields on the bonds that comprise it, where the weights are the proportion invested in each bond.

This is illustrated in Table 21.2. The yield to maturity is calculated for each of the three bonds as well for portfolios of the bonds. In addition, a weighted yield to maturity is calculated where the weights are the proportion invested in each bond. For example, the weights for $A + C$ are

$$\frac{100}{192} \text{ for } A \text{ and } \frac{92}{192} \text{ for } C$$

Note that the weighted average yield to maturity is not the yield to maturity when it is calculated using the cash flows on the portfolio as a whole. Yields are not additive. Investment professionals and some academics often talk about a yield pickup swap as a trading strategy. A yield pickup swap is trading one bond for another bond with a higher yield. Because the yield on a portfolio is not additive, a yield pickup swap can actually lower the yield on a portfolio. Finally, the yield to maturity is not generally the expected return on the bond if the bond is sold before the maturity.

Before leaving this section, we should point out that price bears an inverse relationship with yield to maturity. Figure 21.1 plots the price of a bond for yields ranging from 4%

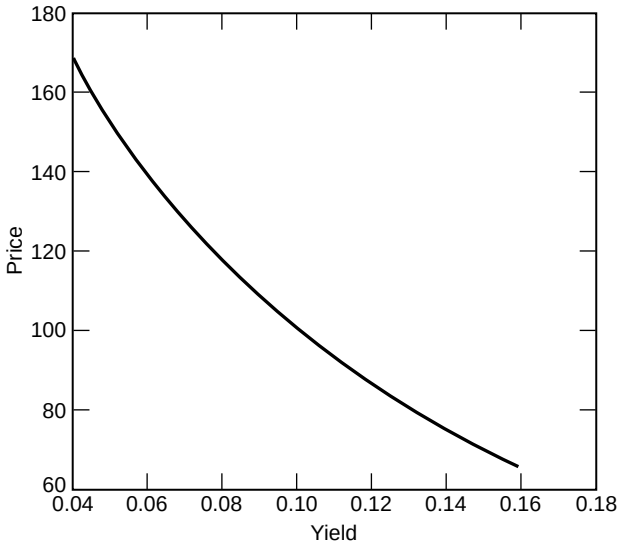


Figure 21.1 Graph of yield versus price.

to 16%. The bond is a 10-year bond with an interest rate of 10% and paying interest semiannually. Note as yield increases, the price declines. Furthermore, the plot has curvature and is not a straight line. This curvature is known as convexity and will be discussed in the next chapter.

A second type of rate frequently quoted in the financial community is current yield. Current yield is simply the annual coupon payment divided by the price. If a bond pays $50 semiannually and costs $800, its current yield is 12.5%. This is determined by 100/800 = 12.5%. Current yield is the "interest rate" normally quoted in the financial press. It has very limited usefulness. Current yield is not the expected return over the year, nor is it the return if the bond is held to maturity. For example, the current yield on a bond that does not pay interest (a zero coupon bond) is zero. An investor who selects investment on the basis of current yield will reject bonds with low coupons but large return in the form of capital gains.

A third type of rate of interest is the spot rate. Spot interest rates are yields to maturity on loans or bonds that pay only one cash flow to the investor. A bond with only one cash flow paid the investor is called a pure discount bond or a zero coupon bond. Spot rates have special importance in bond valuation. As we show in the next section, unless bonds are priced at a price equal to the present value of their cash flows discounted at the spot rate, profitable swaps will exist.

A bond that involves an investment of $970.87 and returns a principal of $1,000 in six months is a six-month pure discount bond. The return on such a bond is the six-month spot rate.

Spot rates are usually calculated for six-month intervals and then annualized by doubling the six-month rate. In what follows, subscripts will designate time, and time will be in six-month intervals. More specifically, in what follows, 0 will designate today, 1 six months from now, 2 12 months from now, and so on.

Defining S_{01} as the annualized spot rate between zero and one,

$$970.87 = \frac{1000}{\left(1 + \frac{S_{01}}{2}\right)} \qquad S_{01} = 6\%$$

Table 21.3 presents a number of other examples of spot rates. Each bond is assumed to cost the amount shown. The cash flows associated with each bond are as indicated. Note that these cash flows involve only a principal payment. Until the early 1980s the only pure discount bonds were those issued by the U.S. government with maturities of one year or less (Treasury bills). As a result, more complex techniques involving the inference of spot rates from coupon-paying bonds were necessary to estimate longer-term spot rates. The techniques used in this calculation are discussed in Appendix B.

In the early 1980s, corporations started to issue pure discount instruments with longer maturities, and brokerage firms put together packages of coupon bonds and sold off each year's payment separately, thereby creating pure discount bonds. These bonds were called *stripped coupon* bonds.

A fourth type of interest rate is the forward rate. Forward rates are interest rates on bonds where the date the commitment is made and the date the money is loaned are different. If a commitment is made now on a one-year loan to commence in six months, then the interest rate on this loan is a forward rate. For example, assume $924.56 is to be lent in 6 months and $1,000 is to be repaid in 18 months; the rate of interest on this loan is a forward rate. As with spot rates, forward rates are estimated for six-month intervals, and then the six-month rate is doubled to annualize.

Table 21.3 Cash Flow with Pure Discount Bonds

		Cash Flows of Pure Discount Bond Cash Inflows						Determination of Spot Rate	
Maturity in Half Years	Cost	1	2	3	4	5	6	Calculation	6-Month Spot Rate (Annualized)
1	970.87	1000						$\left(1+\frac{S_{01}}{2}\right)^1=\frac{1000}{920.87}$	$S_{01}=6\%$
2	933.51	0	1000					$\left(1+\frac{S_{02}}{2}\right)^2=\frac{1000}{933.51}$	$S_{02}=7\%$
3	889.00	0	0	1000				$\left(1+\frac{S_{03}}{2}\right)^3=\frac{1000}{889.00}$	$S_{03}=8\%$
4	838.56	0	0	0	1000			$\left(1+\frac{S_{04}}{2}\right)^4=\frac{1000}{838.56}$	$S_{04}=9\%$
5	783.53	0	0	0	0	1000		$\left(1+\frac{S_{05}}{2}\right)^5=\frac{1000}{783.53}$	$S_{05}=10\%$
6	725.25	0	0	0	0	0	1000	$\left(1+\frac{S_{06}}{2}\right)^6=\frac{1000}{725.25}$	$S_{06}=11\%$

Thus the forward rate from 6 months to 18 months (f_{13}) is calculated by

$$\left(1+\frac{f_{13}}{2}\right)^2=\frac{1000}{924.56}$$

As a second example, consider the interest rate on a two-year loan to be made in one year. Because periods are six-month periods, the loan started at period 2 goes to period 6. If \$845.80 is lent at 2 and \$1,000 is repaid at 6, the annualized forward rate is

$$\left(1+\frac{f_{26}}{2}\right)^4=\frac{1000}{(845.80)}$$

or

$$f_{26}=8\%$$

Forward rates and spots have a very specific relationship. Consider an investor wishing to hold money for two periods. The investor could buy a two-period pure discount instrument. The ending value per \$1 invested would be

$$\$1\left(1+\frac{S_{02}}{2}\right)^2$$

Alternatively, the investor could buy a one-period pure discount instrument and simultaneously agree to invest the proceeds at one at the forward rate from one to two. The ending value per \$1 invested would be

$$\$1\left(1+\frac{S_{01}}{2}\right)\left(1+\frac{f_{12}}{2}\right)$$

Because the forward rate is known at time zero and the commitment is made at zero, the investor can analyze which is better at zero. For there not to be arbitrage opportunities

(buying the more attractive and financing this by issuing the less attractive), the return must be the same, or

$$\left(1+\frac{S_{02}}{2}\right)^2=\left(1+\frac{S_{01}}{2}\right)\left(1+\frac{f_{12}}{2}\right)$$

Therefore

$$\left(1+\frac{f_{12}}{2}\right)=\frac{\left(1+\frac{S_{02}}{2}\right)^2}{\left(1+\frac{S_{01}}{2}\right)}$$

The return equivalency is an application of the law of one price. Similarly, an investor with a three-period horizon could hold a three-period spot or buy a two-period spot and simultaneously enter into a forward commitment from two to three. For there to be no arbitrage the return must be the same, or

$$\left(1+\frac{S_{03}}{2}\right)^3=\left(1+\frac{S_{02}}{2}\right)^2\left(1+\frac{f_{23}}{2}\right)$$

Thus

$$\left(1+\frac{f_{23}}{2}\right)=\frac{\left(1+\frac{S_{03}}{2}\right)^3}{\left(1+\frac{S_{02}}{2}\right)^2}$$

As a further example, consider the spot rates shown in Table 21.3 for period 1 and period 2:

$$S_{01}=6\%$$
$$S_{02}=7\%$$

These can be used to determine the forward rate from period 1 to 2. Thus

$$\left(1+\frac{f_{12}}{2}\right)=\frac{\left(1+\frac{0.07}{2}\right)^2}{\left(1+\frac{0.06}{2}\right)}$$
$$f_{12}=8\%$$

A number of additional examples are shown in Table 21.4. Having examined alternative definitions of rates on bonds, it is time to explain the key role that spot rates play in the pricing of bonds.

BOND PRICES AND SPOT RATES

Table 21.5 shows the cash flows associated with three different bonds. The bonds have cash flows in two periods. The cash flows from bond *A* can be reproduced by taking $\frac{22}{21}$ of bond *B* and $\frac{1}{21}$ of bond *C*. Thus an investor who desires the cash flow pattern of bond *A* can either purchase bond *A* directly or $\frac{22}{21}$ of bond *B* and $\frac{1}{21}$ of bond *C*. Do not be disturbed that the weights do not add up to 1. These are not portfolio weights representing the proportion of the money placed in each asset; rather, they represent how much of *B* and *C* must be purchased to duplicate the cash flows of bond *A*. If the fractions are

Table 21.4 Determination of Forward Rates

Maturity	6-Month Spot Rate	Forward Calculation	Forward Rate (Annualized)
1	3%		
2	3.5%	$\left(1+\frac{f_{12}}{2}\right)=\frac{(1.035)^2}{(1.03)}$	$f_{12}=8\%$
3	4.0%	$\left(1+\frac{f_{23}}{2}\right)=\frac{(1.04)^3}{(1.035)^2}$	$f_{23}=10.01\%$
4	4.5%	$\left(1+\frac{f_{34}}{2}\right)=\frac{(1.045)^4}{(1.04)^3}$	$f_{34}=12.03\%$
5	5.0%	$\left(1+\frac{f_{45}}{2}\right)=\frac{(1.05)^5}{(1.045)^4}$	$f_{45}=14.05\%$
6	5.5%	$\left(1+\frac{f_{56}}{2}\right)=\frac{(1.055)^6}{(1.05)^5}$	$f_{56}=16.07\%$

bothersome, the equivalent transaction is the purchase of 22 bond *B*s and 1 bond *C* to duplicate the cash flow of 21 bond *A*s.

Assume that bond *A* is more expensive than the corresponding portfolio of bonds *B* and *C*. Then an investor wishing to hold bond *A* could purchase the equivalent more cheaply by buying a combination of bonds *B* and *C*. Similarly, an investor holding bond *A* could sell bond *A* and replace it with $\frac{22}{21}$ of bond *B* and $\frac{1}{21}$ of bond *C*. The portfolio would still have the same cash flow, but the investor would obtain an immediate riskless profit equal to the difference in price of bond *A* and the price of the portfolio of bonds *B* and *C* less transaction costs.

A similar argument can be made if bond *A* is cheaper. In this case, any investor holding bonds *B* and *C* could replace an appropriate mixture of them with bond *A*, maintain the same cash flows, and obtain an immediate riskless profit. The belief that the price of *A* should be equal to the price of an appropriate mixture of *B* and *C* is an application of the law of one price.

The law of one price states that two identical items should sell at the same price. In this case, the identical items are the cash flows of bond *A* and the cash flows from the portfolio of $\frac{22}{21}$ of bond *B* and $\frac{1}{21}$ of bond *C*. If these items do not sell at the same price, then everyone interested in the bonds will buy the cheaper, or anyone holding the more expensive bond will swap the more expensive for the cheaper, until they are the same price.

The law of one price has an important implication for bond pricing. It implies that if bonds *A*, *B*, and *C* are of identical risk, such as all government bonds, then alternative cash flows arising in the same period must be discounted at an identical rate. This does not imply that the same rate is used each period, just that all cash flows that occur in the same

Table 21.5 Cash Flows Associated with Three Different Bonds

		Cash Inflows	
Bond	Price	1	2
A	P_A	10	110
B	P_B	5	105
C	P_C	100	0

period must be discounted at an identical rate. We can demonstrate why this is true with an example: if

1. $S_{01} = 6\%$

2. $S_{02} = 7\%$

then

$$P_A = \frac{10}{(1+0.06/2)} + \frac{110}{(1+0.07/2)^2} = \$112.39$$

$$P_B = \frac{5}{(1+0.06/2)} + \frac{105}{(1+0.07/2)^2} = \$102.87$$

$$P_C = \frac{100}{(1+0.06/2)} = \$97.09$$

With these prices, the price of bond *A* is equal to the sum of $\frac{22}{21}$ of the price of bond *B* and $\frac{1}{21}$ of the price of bond *C*: $\frac{22}{21}$ (\$102.87) + $\frac{1}{21}$ (\$97.09) = \$112.42. If the discount rate for the cash flows of any of the three bonds is different, then the price of bond *A* is not the same as the price of the portfolio and the law of one price is violated. For example, if the first-period cash flow for bond *B* is discounted at 8% annually or 4% semiannually, its price is \$102, and the price of the portfolio is less expensive than bond *A*. This general principle has to hold for all bonds, including pure discount bonds. Thus the rate used to discount the cash flows is the spot rate. In summary, either bonds are priced so that their price is equal to the present value of their cash flows discounted at the spot rates, or the law of one price is violated and swap opportunities are available. As discussed previously, forward rates can be derived from spot rates; therefore forward rates can be used equally well to determine bond prices.

DETERMINING SPOT RATES

More details on the techniques for determining spot rates or equivalent discount functions are discussed in Appendix B and the associated references. However, because spot rates and discount functions play such an important role in bond pricing, some understanding of how they are obtained is useful. To illustrate how spot rates are estimated, assume we observe the following two bond prices and cash flows:

		Cash Flows	
Bond	Price	1	2
A	\$100	106	
B	\$ 96.54	6	106

Bond A is a one-period pure discount instrument. Thus the one-period spot rate can be determined directly:

$$100 = \frac{106}{1+\frac{S_{01}}{2}}$$

By inspection, S_{01} is 12%. Having calculated the one-period spot rate, the two-period spot rate can be determined. In the prior section we learned that the price of a bond was the cash flows brought back to present at the spot rate. In symbols this is

$$96.54 = \frac{6}{\left(1+\frac{S_{01}}{2}\right)} + \frac{106}{\left(1+\frac{S_{02}}{2}\right)^2}$$

Because the one-period spot rate was calculated using the one-period bond (bond A), the value of the one-period spot rate can be substituted into the equation.

Substituting 12% for S_{01} leaves S_{02} as the only unknown, and the equation becomes

$$96.54 = \frac{6}{(1.06)} + \frac{106}{\left(1+\frac{S_{02}}{2}\right)^2}$$

Solving for S_{02}, we obtain $S_{02} = 16\%$. Clearly a three-period bond could be used to determine S_{03} and so forth until all spots were determined. There are generally a number of bonds with the same pattern of cash flows, and each of these could be used to derive the spot rates. Because in equilibrium the price of each of these is determined by the same spot rates, it shouldn't matter which was used. In practice, it does matter, and very different spot rates would be estimated, depending on which set of bonds was used to estimate the spots. Some of the reasons for these differences are that bonds differ in tax treatment and callability features. These differences could and should be specifically taken into account. Even without differences in bond characteristics, however, it would matter which bonds were utilized to calculate discount functions and spot rates because of bid–ask spreads and because the prices used in the calculation are often from trades that occurred at different points in time (nonsynchronous trades). For example, a bond dealer might be willing to pay $85\frac{1}{4}$ for a bond but would require $85\frac{1}{2}$ to sell it. Depending on whether the trade was a purchase or sale, the trade price could be $85\frac{1}{4}$ or $85\frac{1}{2}$. Small differences such as this and nonsynchronous trades can result in large differences in estimated discount functions.[5]

What is desired, then, is an average estimate of the spot rates. Multiple regression is an averaging technique. For ease of discussion, we will work with discount functions where

$$d_t = \frac{1}{\left(1+\frac{S_{0t}}{2}\right)^t}$$

[5]In the last few years a large number of pure discount or zero coupon bonds have been introduced in the market. Many of these are stripped governments. Each could be used to easily estimate spot rates. Several factors affect the accuracy and usefulness of direct observation. First, a number of the zero coupon bonds are inactive so that current prices may not exist. Of more importance, the sum of the prices of strips generally is more than the price of the original bond. A higher price implies that spot rates calculated from strips are less than rates used to value coupon bonds. Differences of 1/2% for strips with a long maturity are not uncommon.

Obviously the differences just discussed violate the law of one price because the aggregate value of the strips is higher than the value of the bond or bonds that were stripped. The relevant question is, How can this exist in equilibrium? The action that would force identical prices is to buy the least expensive (the bond or bonds being stripped) and sell the more expensive zero coupon strips. However, individual investors cannot issue strips. A creator of strips must be able to obtain the trust of investors, and this requires a large brokerage firm such as Salomon or Merrill Lynch. The difference in the aggregate value of the strips and the cost of the bonds being stripped is their profit. Competition will narrow the difference but not eliminate the difference entirely because the brokerage firms will only issue strips if there is a profit to be made. Thus, in equilibrium, the aggregate value of the strips can be different from the value of the underlying securities.

obviously knowing d_t allows calculation of spot rates. The price of bond i can be expressed as the present value of the cash inflows or

$$P_i = d_1 C_i(1) + d_2 C_i(2) + d_3 C_i(3)$$

where

P_i is the price of bond i

$C_i(t)$ is the cash flow on bond i in period t

d_t are the discount functions

We expect to have many bonds with the same cash flow patterns. As discussed previously, they will have prices different from this equation because of nonsynchronous trading, bid–ask spreads, and possibly nonequilibrium prices as well as differences in bond characteristics. To account for these differences, a random error term (e_i) is added:

$$P_i = d_1 C_i(1) + d_2 C_i(2) + d_3 C_i(3) + e_i \tag{21.1}$$

For any bond, the price (P_i) and cash flows ($C_i(t)$) would be known. The discount functions are analogous to coefficients in a normal regression. The data used are the prices and cash flows on a sample group of bonds. The discount factors are outputs of the normal regression. Thus Equation (21.1) could be used to estimate discount functions and hence spot rates. In practice, terms to account for tax considerations and callability are usually added to the equation. Because most bonds do not pay interest on the same dates, the procedures used by many firms for estimating discount functions are somewhat more complicated. These are discussed in Appendix B and the associated references.

Spot rate estimation is important and is the starting point for most organizations involved in bond management. Many organizations simply use spot rates to understand the returns in the market for different holding periods. Others use estimated spots to price strips or zero coupon debt. The organization estimates the spot rates and then prices zeros to yield a rate so many hundredths of a percent different from the spot. A third use for spots is finding mispriced bonds. Those bonds with model prices [as determined by Equation (21.1)] that are very different from actual price are examined to see if there is an explanation for the mispricing. If there is not, these bonds become candidates for purchase or sale.

A final use of estimated spots is in pricing private placements. A large portion of the debt market involves loans from financial intermediaries such as banks or insurance companies to corporations. One of the advantages of private placements relative to the public market is that unusual cash flow patterns can be set (involving uneven interest and principal payments) to better match the corporation's cash generation pattern. These unusual patterns cannot be priced relative to the public market, because public counterparts do not exist. Estimated spot rates are used to price private placements with unusual cash flow patterns.

THE DETERMINANTS OF BOND PRICES

Bonds can differ in a number of respects. These differences affect bond prices, spot rate, yields to maturity, the expected return in the next period, and the risk associated with next period's return. Standard bond theory deals with the determination of the yield to maturity or price. The yields to maturity on bonds differ for a number of reasons. Among the more important are the following:

1. the length of time before the bond matures
2. the risk of not receiving coupon and principal payments
3. the tax status of the cash flows

4. the existence of provisions that allow the corporation or government to redeem the debt before maturity
5. the amount of the coupon

Term to Maturity and Term Structure Theory

To gain insight into the effect of maturity on the yield or price of a bond, it is necessary to understand the relationship between yield and time. This relationship is usually called the *term structure*. More precisely, the theory of the term structure of interest rates deals with why pure discount bonds of different maturities have different yields to maturity.[6] In the last section, it was pointed out that spot rates are equivalent to the yield to maturity on pure discount instruments. Thus term structure theory could be described equally well as dealing with the determination of spot rates.

In analyzing the effect of maturity on yield, all other influences are held constant. Pure discount instruments are chosen to eliminate the effect of coupon payments. In addition, most analysis is done using government bonds without early redemption features. Therefore bonds of different maturities are similar with respect to risk, tax liabilities, and redemption possibilities.

Figures 21.2 and 21.3 depict two different yield curves. Figure 21.2 shows a yield curve where the yield to maturity declines as maturity increases. In Figure 21.3 the yield curve has a more normal upward slope. Term structure theory deals with why we observe these different shapes. In the next sections we discuss four different explanations.

Segmented Market Theory Segmented market theory has its origin in the observation that many investors and issuers of debt seem to have a strong preference for debt of a certain maturity. Furthermore, they seem to be insensitive to differentials in yields between debt of this maturity and debt of a different maturity.

Consider first debt with a long maturity. Let us examine the problem of maturity selection from the viewpoint of an insurance company. Life insurance companies offer insurance policies that are unlikely to require any payment for a long time. An insurance policy issued to a 25-year-old individual may involve 25 or more years before the company anticipates having to make a payment. The size of the premium payments is determined in part by the anticipated interest rate. If the insurance company invests in a long-term bond, the interest earned

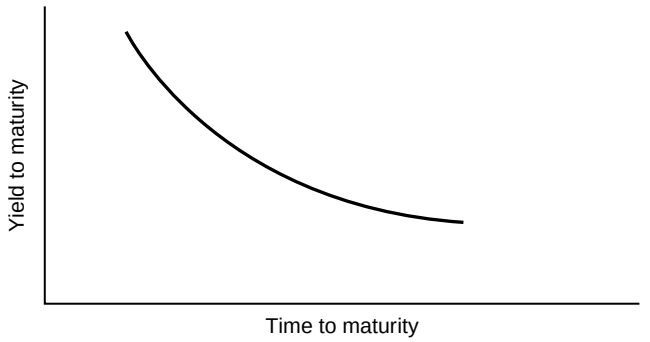


Figure 21.2 Possible term structure.

[6]Term structure theory is often incorrectly defined as explaining why coupon-paying bonds of different maturities have different yields to maturity.

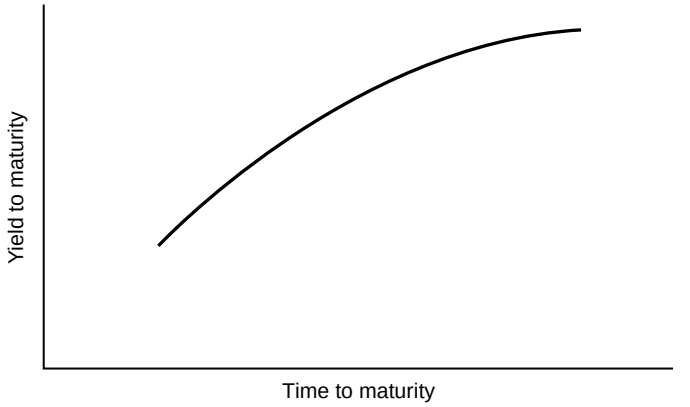


Figure 21.3 Possible term structure.

on the bond is known, and if it exceeds what was promised on the insurance contract, it substantially reduces the insurance company's risk. There is still some risk because the coupon payments will have to be reinvested at some future unknown rate. However, the principal remains invested at a known rate, which substantially reduces the risk. Alternatively, the insurance company could meet its long-term obligation by buying a sequence of one-year bonds. However, in this case, all earnings beyond the first year are unknown. If interest rates decline below what was anticipated in the insurance contract, the company may have difficulty meeting its obligations. Not only is there uncertainty associated with the rate that will be earned on the investment of the coupon payments, there is also uncertainty about the rate earned on the principal. Consequently, many insurance companies invest in long-term bonds even when short-term rates are considerably higher than long-term rates.

Let us examine the maturity selection problem from the viewpoint of the issuers of long-term debt. The construction of a manufacturing plant or warehouse or other physical facility can involve a large expenditure of funds for a corporation. These structures are long-lived assets. Corporations normally wish to pay for them over a long period of time. They can achieve this payment pattern by issuing long-term debt. Alternatively, they can issue short-term debt and keep reissuing it for a long period of time. If they issue the long-term debt, their costs are known ahead of time and there is no interest rate risk associated with the investment. This suggests that corporations will generally issue long-term debt to meet these types of obligations.

Similar considerations apply to short-term debt. Corporations have a number of known short-term obligations that occur at fixed intervals: tax payments and wages are two examples. Money is normally put aside to meet these obligations. If the corporation buys pure discount securities maturing exactly on the date the payment is due, they have zero risk concerning the amount of money they will have available. If they buy a longer-term security, the treasurer faces the risk that interest rates will increase, the price of the security will fall, and the amount that will be available to meet the obligations will be less than anticipated. Commercial banks hold a large number of short-term securities. For example, checking accounts make up a large percentage of the liabilities of commercial banks. Commercial banks engage in short-term lending to match the maturity of their assets with the maturity of their debt.

Market segmentation theory argues that investors are sufficiently risk averse that they operate only in their desired maturity spectrum. No yield differential will induce them to change maturities. Thus what determines long-term rates is solely the supply and demand of long-term funds. Similarly, short-term rates are determined only by supply and demand

of short-term funds. People who believe in market segmentation theory examine flows of funds into these market segments to predict changes in the yield curve.

Market segmentation theory is very popular with practitioners. Statements in the popular press often display an implicit belief in the market segmentation theory. The theory is much less popular with academics, who maintain that while there are investors who have strong maturity preferences, there are others who are attracted by relative yields. The effects of segmentation on interest rates will be offset if there are enough such investors.

Pure Expectations Theory The pure expectations theory explains the term structure in terms of expected one-period spot rates. Advocates of the expectations theory believe that the yield on a one-year bond is set so that the return on the one-year bond is the same as the return on a six-month bond plus the expected return on a six-month bond purchased six months hence.

If the expectations theory is correct, then an upward sloping yield curve is an indication that short-term rates are expected to increase. Similarly, a flat yield curve is an indication that short-term rates are expected to remain the same. Finally, a downward sloping yield curve indicates that short-term rates are expected to decline.

The easiest way to understand the expectations theory is to assume that the investors setting prices do not care about risk (are risk neutral). In this case, no matter what their time horizon, they will select the security or securities that give them the highest expected return. This is exactly the opposite of the market segmentation theory.

Consider an investor with a one-year time horizon. Assume that the yield to maturity on a pure discount six-month bond is 10% and on a one-year pure discount bond is 12%. Furthermore, assume that the investor expects the six-month spot rate to be 16% in six months. The one-year investment can be accomplished by holding a one-year bond with earnings per dollar invested, assuming semiannual compounding of

$$\left(1+\frac{0.12}{2}\right)^2-1=1.1236-1=12.4 \quad \text{or } 12.4\%$$

Alternatively, the investor can hold two six-month bonds with expected earnings per dollar invested of

$$\left(1+\frac{0.10}{2}\right)\left(1+\frac{0.16}{2}\right)-1=1.134-1=13.4 \quad \text{or } 13.4\%$$

The 16% is, of course, the expected one-period spot rate six months in the future. Given this combination of observed and expected rates, holding two six-month bonds gives the higher return, and all two-period investors will wish to hold the two one-period bonds.[7]

---

[7]The same choice would be made by investors with six-month horizons. These investors have the choice of buying the six-month bond or the one-year bond and selling it in six months. In six months the one-year bond will have six months remaining in its life. At that point it will have to offer the same yield as a newly issued bond. With semiannual compounding, the one-year bond will pay \$1.1236 at maturity for each dollar invested. For it to have a 16% annual return with six months left before it matures, its price per dollar invested must be

$$\frac{1.1236}{\left(1+\frac{0.16}{2}\right)}=1.0404$$

If, instead, the investor buys a six-month bond, its value will be

$$\$1\left(1+\frac{0.10}{2}\right)=\$1.05$$

Table 21.6 Two Hypothesized Sequences of Expected One-Period Rates

	Upward Yield Curve		Downward Yield Curve	
Period	Expected One-Period Spot Rates	Yield to Maturity	Expected One-Period Spot Rates	Yield to Maturity
1	10	10.0	10	10.0
2	11	10.5	9	9.5
3	12	11.0	8	9.0
4	13	11.5	7	8.5
5	14	12.0	6	8.0
6	15	12.5	5	7.5
7	16	13.0	5	7.1
8	16	13.4	5	6.9
9	16	13.7	5	6.7
10	16	13.9	5	6.5

We have analyzed the return for investors with two-period horizons. The same results apply to investors with any other horizon. Given this universal preference, prices should adjust until the expected return from holding a one-year bond is exactly the same as the expected return from holding two six-month bonds.

Under the expectations theory the yield curve can be derived directly from a series of expected one-period spot rates. Table 21.6 shows two hypothesized sequences of expected one-period rates. One of these sequences produces an upward sloping yield curve, whereas the other sequence produces a downward sloping yield curve.

Let us examine an example of the calculations. Under the expectations theory investing in a two-period bond and earning the spot rate from 0 to 2 must produce the same expected return as investing in two one-period bonds earning the spot rate from 0 to 1 and the expected spot rate from 1 to 2. Thus, in Table 21.6, S_{02} is calculated from

$$\left(1+S_{02}/2\right)^2 = \left(1+0.10/2\right)\left(1+0.11/2\right)$$
$$S_{02} = 10.5\%$$

Similarly,

$$\left(1+S_{03}/2\right)^3 = \left(1+0.10/2\right)\left(1+0.11/2\right)\left(1+0.12/2\right)$$
$$S_{03} = 11\%$$

Not only can the yield curve be derived from the expected spot rates, but under the expectations theory, the market's belief about future one-period rates can easily be derived from an observed yield curve.

It is important to keep in mind the distinction between the six-month rate expected to prevail six months from now ($\overline{S}_{12}$) and the forward rate, f_{12}. The expectations theory simply states that the two must be equal.[8] In the next two sections we examine alternative theories under which they are no longer equal.

Liquidity Premium Theory Liquidity premium theory is also based on investors analyzing the returns from holding bonds of varying maturities. However, unlike expectations

[8] S_{12} is the spot rate that is expected to prevail at time 1. The expectation is as of time 0.

theory, liquidity premium theory assumes investors must be offered a higher expected return to hold a bond with a horizon different from their preferred horizon. Furthermore, it is assumed that there is a shortage of longer-term investors so that extra return must be offered on long-term bonds to induce investors to hold them.

In the prior example, we considered investors with one- and two-period time horizons. We assumed the one-period rate was 10% and the one-period rate that was expected to prevail one period hence was 16%. Under the expectations theory, the two-period rate would be 13%. With the liquidity premium theory, this rate would have to be higher. The assumption is that there is an excess of investors with short-term horizons.

These investors have a choice of holding a six-month bond or of holding a one-year bond and selling it in six months. The investment in the one-year bond involves risk to the six-month investor. To induce some six-month investors to hold one-year bonds, a premium will have to be offered. Thus the return from holding a one-year bond will be above the expected return from holding two six-month bonds.

For an investor with a six-month horizon, a bond with a maturity longer than one year is even riskier than a one-year bond. Thus an even larger premium would be required on three- and four-period bonds. If the market is dominated by short-term investors, then the longer-term bonds will require larger premiums. This is the basic idea behind liquidity premium theory. Note that if the liquidity premium theory holds, an investor with a long-term horizon can hold a bond matching his horizon and earn the liquidity premium. Thus such an investor earns an extra return without any extra risk.

In Table 21.7 we have taken the returns from Table 21.6 and added the liquidity premium. These are then used to construct a term structure. For example, for period 3, the yield to maturity was calculated by solving for S_{03}, where

$$\left(1+\frac{S_{03}}{2}\right)^3=\left(1+\frac{0.10}{2}\right)\left(1+\frac{0.11+0.002}{2}\right)\left(1+\frac{0.12+0.004}{2}\right)$$

$$S_{03}=11.2\%$$

Liquidity premium theory modifies the conclusions drawn in the prior section concerning the shape of the yield curve and the implied one-period rates in future periods. If expectations are for an unchanged one-period rate, then the presence of a liquidity premium imparts an upward sloping shape to the yield curve. Even if expectations are for a

Table 21.7 Yield Curve with a Liquidity Premium (Expressed in Percentage)

	Upward Sloping Yield Curve			Downward Sloping Yield Curve		
Period	Expected One-Period Spot Rate	Liquidity Premium	Yield to Maturity	Expected One-Period Spot Rates	Liquidity Premium	Yield to Maturity
1	10	0	10.00	10	0	10.00
2	11	0.2	10.60	9	0.2	9.60
3	12	0.4	11.20	8	0.4	9.20
4	13	0.6	11.80	7	0.6	8.80
5	14	0.8	12.39	6	0.8	8.40
6	15	1.0	12.99	5	1.0	8.00
7	16	1.2	13.59	5	1.2	7.74
8	16	1.4	14.06	5	1.6	7.57
9	16	1.6	14.45	5	2.0	7.46
10	16	1.8	14.79	5	2.4	7.40

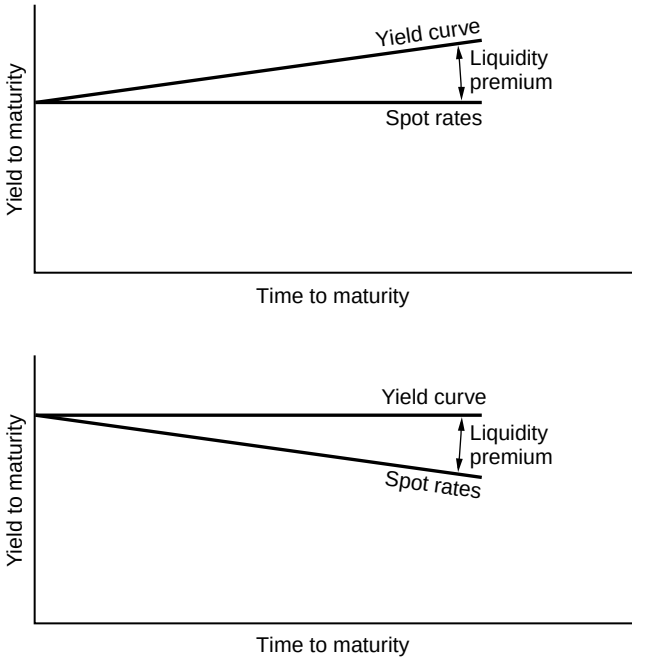


Figure 21.4 Yield curves with liquidity premiums.

declining series of one-period rates, it is still possible to observe an upward sloping yield curve. This would occur if the risk premiums were sufficiently large to overcome the expectations of a decline in one-period rates. Thus an upward sloping yield curve would be consistent with any pattern of expectations concerning one-period rates. A flat or downward sloping yield curve is only consistent with a decrease in one-period rates. Figure 21.4 depicts two yield curves and the associated liquidity premiums.

Preferred Habitat Preferred habitat theory rests on the premise that investors who match the life of their assets with the life of their liabilities are in the lowest risk position. Matching the life of the assets and liabilities is their preferred position. If there is sufficient extra return to be earned on assets of other lives, they will adjust their position to include more of these higher-yielding assets.

If this theory is correct, premiums will exist for maturities where there is insufficient demand. These premiums are necessary to induce investors to leave their preferred habitat. If there are a large number of firms issuing long-term debt relative to the number of investors interested in long-term debt, a premium will have to be offered on long-term debt. If many firms and institutions wish to issue short-term debt and there are few investors who wish to invest short terms, a premium will have to be offered on short-term debt.

What is meant by a premium? For simplicity consider two periods. Let S_{01} be the spot interest rate in the first period and $\overline{S}_{12}$ be the expected one-period spot rate in the second. If the expectations theory holds, the two-period rate expressed as a rate per period is

$$\left(1+\frac{S_{02}}{2}\right)^2=\left(1+\frac{S_{01}}{2}\right)\left(1+\frac{\overline{S}_{12}}{2}\right)$$

Assume that there is a surplus of short-term investors and therefore an extra return is necessary to induce investors to hold the two-period bond. If P is the size of the premium, then

$$\left(1+\frac{S_{02}}{2}\right)^2=\left(1+\frac{S_{01}}{2}\right)\left(1+\frac{\bar{S}_{12}}{2}+\frac{P}{2}\right)$$
$$P>0$$

In this case, preferred habitat theory would result in a set of spot rates that could have been derived equally well from the liquidity premium theory. If, on the other hand, there is a need to move investors to the short term, holding the two-period bond will be less profitable than holding two one-period bonds, or

$$\left(1+\frac{S_{02}}{2}\right)^2=\left(1+\frac{S_{01}}{2}\right)\left(1+\frac{\bar{S}_{12}}{2}+\frac{P}{2}\right)$$

with

$$P<0$$

With the preferred habitat theory, the premiums can be positive or negative. Without an idea of the sign and size of the premiums, nothing can be concluded about future one-period rates from observing the yield curve.

Term Structure and Coupon Bonds In the last section we examined the term structure for pure discount bonds. We will now examine the term structure for coupon-paying bonds. A coupon-paying bond can be considered a portfolio of pure discount bonds. Consider a three-period bond with a coupon of \$75 and a principal repayment of \$1,000. Its price is calculated as follows:

$$\text{Price}=\frac{75}{\left(1+\frac{S_{01}}{2}\right)}+\frac{75}{\left(1+\frac{S_{02}}{2}\right)^2}+\frac{1{,}075}{\left(1+\frac{S_{03}}{2}\right)^3} \tag{21.2}$$

This bond can be viewed as one bond or as a portfolio of three bonds—one-period, two-period, and three-period pure discount bonds paying \$75, \$75, and \$1,075, respectively. The price on this portfolio is given by Equation (21.2).

The price of the portfolio is, of course, the same as the price of the bond. The yield to maturity on the bond lies between the spot rates. Let us examine what this implies for yield curves of coupon bonds relative to yield curves of pure discount bonds. Consider a downward sloping yield curve. The spot rates associated with the earlier coupon payments are higher than the spot rate associated with the final maturity. Because the yield to maturity lies between these rates, the yield to maturity on the coupon bond lies above the spot rate associated with the final payment (see Figure 21.5). The higher the coupon payments, the greater the importance of earlier payments relative to the last payment and the more important the influence of earlier spot rates on the yield to maturity. Thus the higher the coupon payment, the greater the difference between the yield to maturity on the coupon-paying bond and the spot rate on the final payment.

Figure 21.5 shows the plot of yield to maturity on coupon bonds compared to pure discount bonds. As just discussed, the greater the coupon, the greater the difference between yields to maturity and the spot rate of the final payment.

If the yield curve is upward sloping, then the yield to maturity on coupon bonds lies below the yield to maturity on discount bonds. The larger the coupon, the greater the difference

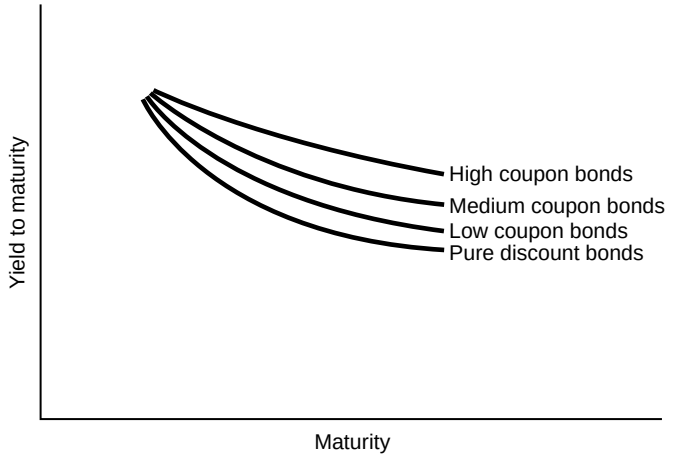


Figure 21.5 Possible term structure curves.

between the yield on the coupon and noncoupon debt. Figure 21.6 plots the yield to maturity on bonds with various coupons with upward sloping yield curves.

A number of organizations examine yield curves on coupon paying debt. Pure discount debt for government bonds did not exist at all for bonds with maturities over one year until the 1980s. When pure discount debt for longer maturities was first offered, it was created by brokerage firms removing coupons from coupon bonds and selling them off separately. These instruments are not quite equivalent to pure discount government bonds, since they may be less marketable than when the government originally issued them, and there is some risk of the brokerage firm defaulting. Furthermore, even now there are not enough of them to allow accurate estimation of the yield curve. Most firms plot yield curves of coupon paying debt rather than go through the process of estimating the yield curve for pure discount debt using techniques discussed in the appendix at the end of the chapter. Examining Figures 21.5 and 21.6 shows that the general shape of the yield curve is preserved if the

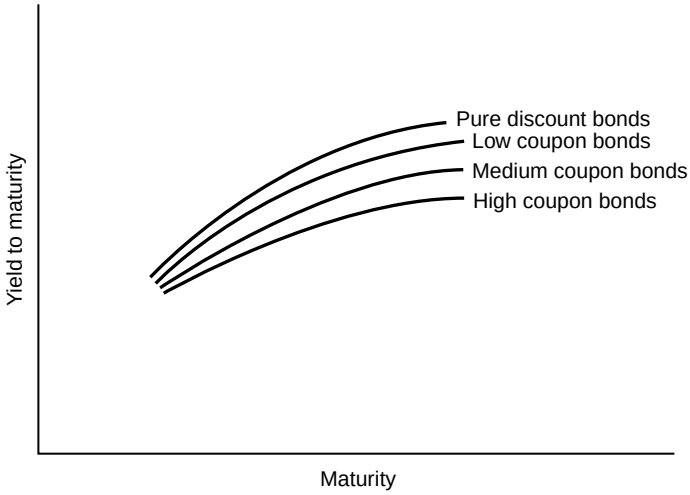


Figure 21.6 Possible term structure curves.

coupon rate on bonds of varying maturities is the same. The problem is that they are not the same. Most bonds with intermediate maturity are long-term bonds that were issued several years before. For example, a bond with a 7-year maturity might be a 30-year bond issued 23 years ago. Interest rates change dramatically over time. Thus the coupon rate on bonds of different maturities is likely to be very different. A yield curve drawn from coupon-paying bonds is likely to be a mixture of the yield curves shown in Figures 21.5 and 21.6. In this case, even the shape need not be preserved.

Organizations examine yield curves for investment decisions and for determining interest rates to be offered their customers. Using coupon bonds can lead to very misleading yield curves and incorrect decisions.

Summary of the Term Structure of Interest Rates We have shown how spot rates can be used to arrive at the correct price of any bond. To estimate spot rates, one should use the methodology outlined in the appendix at the end of this chapter. Spot rates are determined by current one-period rates, expectations about future one-period rates, theories of institutional behavior, and risk preferences. Although we have not attempted to find a categorical answer to which of these term structure theories is correct, we have provided you with enough information about the contrasting theories to give insight into the term structure of interest rates.

Default Risk

Unlike government bonds, for corporate bonds and municipal bonds, there is a risk that the coupon or principal payments will not be met. For these bonds it is necessary to make a distinction between promised return and expected return. A bond could promise a return of 12%, but if there were some probability that the principal or coupon might not be paid, its expected return could be 10%. In addition, because there is risk associated with these bonds, investors should require that the expected return be greater than the return on a similar bond that is default free. These concepts are illustrated in Table 21.8.

We have referred to the difference between the promised return and the expected return as the default premium. The difference between the expected return and the return on a default-free instrument is the risk premium. The investor requires this extra return because of the chance that a particular bond selected may default, resulting in a very poor and probably negative return.

Three large investment services estimate the likelihood of default for most corporate bonds: Moody's, Standard and Poor's, and Fitch. The estimates from Moody's and from Standard and Poor's are widely available. Their services are similar in that they classify bonds by likelihood of loss. Likelihood of loss includes both the probability of a missed, delayed, or partial payment and the size of the loss if a loss occurs. For example, consider two bonds with the same probability of a missed principal payment. If one of them has significant odds of paying a

Table 21.8 Components of Interest Rates on Corporate Bonds

2%	Default premium	12% Total return
1%	Risk premium	
9%	Return on default-free bonds	

Table 21.9 Key to Moody's Corporate Ratings

Aaa	Bonds that are rated Aaa are judged to be of the best quality. They carry the smallest degree of investment risk and are generally referred to as "gilt edge." Interest payments are protected by a large or by an exceptionally stable margin and principal is secure. While the various protective elements are likely to change, such changes as can be visualized are most unlikely to impair the fundamentally strong position of such issues.
Aa	Bonds that are rated Aa are judged to be of high quality by all standards. Together with the Aaa group they comprise what are generally known as high-grade bonds. They are rated lower than the best bonds because margins of protection may not be as large as in Aaa securities or fluctuation of protective elements may be of greater amplitude or there may be other elements present that make the long-term risks appear somewhat larger than in Aaa securities.
A	Bonds that are rated A possess many favorable investment attributes and are to be considered as upper medium-grade obligations. Factors giving security to principal and interest are considered adequate but elements may be present that suggest a susceptibility to impairment sometime in the future.
Baa	Bonds that are rated Baa are considered as medium-grade obligations (i.e., they are neither highly protected nor poorly secured). Interest payments and principal security appear adequate for the present, but certain protective elements may be lacking or may be characteristically unreliable over any great length of time. Such bonds lack outstanding investment characteristics and in fact have speculative characteristics as well.
Ba	Bonds that are rated Ba are judged to have speculative elements; their future cannot be considered as well assured. Often the protection of interest and principal payments may be very moderate and thereby not well safeguarded during both good and bad times over the future. Uncertainty of position characterizes bonds in this class.
B	Bonds that are rated B generally lack characteristics of the desirable investment. Assurance of interest and principal payments or of maintenance of other terms of the contract over any long period of time may be small.
Caa	Bonds that are rated Caa are of poor standing. Such issues may be in default or there may be present elements of danger with respect to principal or interest.
Ca	Bonds that are rated Ca represent obligations which are speculative in a high degree. Such issues are often in default or have other marked shortcomings.
C	Bonds that are rated C are the lowest-rated class of bonds, and issues so rated can be regarded as having extremely poor prospects of ever attaining any real investment standing.

substantial portion of the principal payment if missed, while the odds are that the other will pay none, then the bond with the higher payment receives the higher rating. Bond rating services divide bonds into discrete classes. Table 21.9 shows Moody's classification of bonds and their discussion of what the various classifications mean.

Many organizations are restricted to buying bonds that have achieved at least a certain rating. These restrictions may be imposed by regulatory authority, by perception of legal requirements of prudent investment, or by organizational policy. In addition, many brokerage firms put together pools of bonds and then issue shares in these pools. These pools are normally restricted to A-rated bonds or better. These restrictions suggest the possibility of a segmented market between higher-rated bonds and lower-rated bonds; however, we know of no conclusive evidence on this issue.

Moody's and Standard and Poor's classifications can be duplicated fairly accurately by utilizing a weighted average of firm characteristics, as follows. A number of firm characteristics are hypothesized as influencing Moody's or Standard and Poor's classifications. These characteristics usually include variables such as the amount of earnings compared to the interest payments, the variability of earnings, the amount of debt in the capital structure, the net worth,

and the amount of short-term assets compared to short-term liabilities. Data on these variables are collected for a number of publicly traded bonds along with the classification of each bond by one of the bond rating services. Mathematical techniques exist for finding the combination of firm variables that best duplicates the classification of the rating agency. The combination is best in the sense that it most accurately reproduces the ratings. Once the best combination is determined, it is then tested using data on other publicly traded bonds to see how well it classifies them. Accurate classification of 70%–80% of the bonds is not uncommon, with most bonds being only one rating away from the published ratings.

Reproducing public ratings is useful in order that bonds not classified by the public rating services can be inexpensively and accurately classified. The most obvious utilization of this system is in classifying private placements. Banks and insurance companies lend money to firms directly. These private placements are usually loans to small- or medium-sized companies that wish to avoid the expenses of issuing publicly traded debt (e.g., SEC registration, brokerage costs). Analysts make judgments concerning the likelihood and size of loss, the appropriate interest rate on the potential loan, and the decision on whether to lend. When individual lending officers are judged in part by the volume of loans they make, they tend to be optimistic about the likelihood of the firm repaying the loan in the future. A scheme that fairly accurately reproduces public ratings is a check on this optimism. These schemes are frequently used to rate all loans under consideration. The analyst is then required to justify any difference in interest rates she wishes to offer compared to what is normal given the rating the bond receives.

Table 21.10 shows the default experience in recent years. Default rate has averaged 3.5%, but there is considerable variability. Another way of examining the default experience is to examine it over the life of the bond. Table 21.11 shows the cumulative default experience for newly issued bonds in each year subsequent to issue. Thus the 60.78% for CCC bonds implies that 60.78% of the bonds rated CCC defaulted in the first 10 years. The default experience over the life of the bond is quite substantial for low-rated bonds outstanding for a number of years. Note also that as discussed earlier, the default experience in the first year tends to be less than in subsequent years.

Tax Effects

The cash flows from certain bonds have a tax advantage. These bonds should sell at a different yield to maturity than bonds without this tax advantage. The most obvious example of such bonds is municipal bonds. The coupon payments from municipal bonds are not subject to federal taxation and usually are not subject to tax in the state where they are issued. Because of the benefits of such favorable tax treatment, the yield to maturity on these bonds is less than the yield to maturity on comparable taxable issues. Generally the yield to maturity is 30%–40% lower on municipal bonds than on similar taxable issues.

The second example of the effect of tax on bonds are the so-called flower bonds. Flower bonds were designated as such at time of issue. These bonds were originally issued at times of relatively low interest rates. Normally they would sell at a value well below face value so that their yield to maturity would be comparable to other bonds. However, they have a unique provision that substantially affects their value. Flower bonds are accepted at face value in payment of estate taxes. Thus a wealthy individual might find it attractive to add flower bonds to her portfolio if an imminent demise were anticipated. Because of this special provision, flower bonds will sell at much higher prices than they otherwise would, leading to lower yields to the investor.

Table 21.10 Historical Default Rates—Straight Bonds Only, 1985–2011 (Dollars in Millions)

Year	Par Value Outstanding (a)($)	Par Value Defaults($)	Default Rates(%)
2011	1,354,649	17,813	1.315
2010	1,221,569	13,809	1.130
2009	1,152,952	123,878	10.744
2008	1,091,000	50,763	4.653
2007	1,075,400	5,473	0.509
2006	993,600	7,559	0.761
2005	1,073,000	36,209	3.375
2004	933,100	11,657	1.249
2003	825,000	38,451	4.661
2002	757,000	96,858	12.795
2001	649,000	63,609	9.801
2000	597,200	30,295	5.073
1999	567,400	23,532	4.147
1998	465,500	7,464	1.603
1997	335,400	4,200	4.252
1996	271,000	3,336	1.231
1995	240,000	4,551	1.896
1994	235,000	3,418	1.454
1993	206,907	2,287	1.105
1992	163,000	5,545	3.402
1991	183,600	18,862	10.273
1990	181,000	18,354	10.140
1989	189,258	8,110	4.285
1988	148,187	3,944	2.662
1987	129,557	7,486	5.778
1986	90,243	3,156	3.497
1985	58,088	992	1.708
Arithmetic Average Default Rate			Standard Deviation
		1985 to 2011	4.093%	3.510%

Notes
(a) As of mid-year.
(b) Weighted by par value of amount outstanding for each year.
Source: Altman and Keuhne (2012).

Table 21.11 Mortality Rates by Original Rating—All Rated Corporate Bonds* (1971–2011)

Rating		Years after Issuance									
		1	2	3	4	5	6	7	8	9	10
AAA		0.00%	0.00%	0.00%	0.00%	0.02%	0.04%	0.05%	0.05%	0.05%	0.05%
AA	Cumulative	0.00%	0.00%	0.25%	0.36%	0.38%	0.40%	0.41%	0.42%	0.45%	0.46%
A	Cumulative	0.01%	0.07%	0.23%	0.40%	0.54%	0.64%	0.68%	0.98%	1.09%	1.15%
BBB	Cumulative	0.38%	2.86%	4.19%	5.20%	5.75%	6.00%	6.28%	6.44%	6.59%	6.93%
BB	Cumulative	1.01%	3.06%	6.89%	8.75%	10.96%	12.27%	13.59%	14.54%	15.82%	18.52%
B	Cumulative	2.96%	10.59%	17.70%	24.22%	28.65%	31.92%	34.41%	35.82%	36.99%	37.51%
CCC	Cumulative	8.30%	19.90%	34.54%	45.24%	47.88%	54.02%	56.53%	58.68%	58.97%	60.78%

*Rated by S&P at Issuance
Based on 2,644 defaulted issues

While flower bonds are the most colorful bond with special tax treatment, the most common type of bond subject to special tax treatment is one with a sufficiently low or high coupon to cause it to sell at a price very different than its face value. For these bonds, capital appreciation or loss is a significant part of the investor's return in addition to interest income. Consider a low coupon bond. The coupon payments are subject to taxation at ordinary income tax rates. Low coupon bonds would have two components to their return: the return from the coupon plus the return from the price appreciation. The total return must be competitive with other bonds of similar characteristics. The portion of return from the price appreciation is taxable as a capital gain. For most investors the capital gain rate is lower than the income tax rate. Thus low coupon bonds have a tax advantage because a portion of their return receives favorable tax treatment. Given this tax advantage, low coupon bonds should and do have a lower (before tax) yield to maturity. McCulloch (1975) has estimated that bonds are priced consistent with investors being in a 20%–30% tax bracket. This means that the after-tax yield on a low coupon versus a normal coupon bond with similar characteristics is the same as if flows were adjusted by assuming a 20%–30% tax bracket.[9]

The tax bracket that is consistent with observed prices is important information to investors. If bonds are priced consistent with a 20%–30% tax bracket, then investors in higher tax brackets will favor low coupon issues, all else held constant. Similarly, tax-exempt investors should primarily be holding the high coupon, high-yield bonds.

Option Features of Bonds

Bonds sometimes contain a feature that constitutes an option for either the issuer of the bond or the holder of the bond. Because the valuation of options is discussed in detail in Chapter 23, we limit our discussion in this chapter to a description of bond features that can be valued as options. Applying the option valuation formula to these features will not be specifically treated, although the option chapter together with the bibliography at the end of this chapter will allow the interested reader to pursue this subject.

The most common option included in bond contracts is the possibility of a call by the issuing firm. The call privilege is the right by the issuing firm to repurchase the bond at a fixed price. The price is generally the par value (face value) of the bond plus a premium (called the call premium). For example, the bond might be callable at par plus 5% of par. Generally the call premium declines over time, making the likelihood of a call higher in later years than in earlier years. For example, the call premium might be 5% in the first year, 4% in the second year, 3% in the third year, and so forth. In addition, it is common to preclude a call for a number of years. The possibility of a call reduces the value of the bond to the investor. An investor can assume that the firm will call at times when the bond without the call feature is worth more than the price at which it is actually called. This difference is a loss to the investor. The value of a comparable noncallable bond will lie above the call price when interest rates decline compared to the original issue price. Thus an investor wishing to lock up high interest rates by buying a bond at a time of high rates might find that he or she earns these rates only for a short time because the bond is called away when rates decline and the proceeds are invested at these lower rates. Many firms calculate return to the first time at which a bond is callable to compare return on callable and noncallable bonds. This procedure makes the unrealistic assumption that firms will call as soon as a bond is callable. This underestimates its value, just as a return to maturity that ignores the possibility of call is an overestimate of

[9]High coupon bonds have a tax disadvantage. Coupon payments are subject to the high ordinary tax rate. The price decline is a long-term loss. However, some of the loss in price may need to be amortized and can be used to reduce the ordinary income.

the expected return over this horizon. The only accurate way to estimate the value of the call is to use the option models discussed in Chapter 23.

Another option associated with bonds is the sinking fund option. Many bond issues require that part of the issue be retired over the life of the bond. For example, bond covenants may require that 5% of the issue be retired at the end of each year over the bond's 10-year life. The corporation has the option of purchasing the bonds directly or of calling the bonds it needs to meet its sinking fund obligation. Obviously it will meet its obligation in the least expensive way. Because the bonds are chosen in a random way, all investors risk having their bonds called to meet the sinking fund obligation. The discussion of the call option is relevant in this case.

A third option found in certain bond contracts is the conversion option. This option benefits the bondholders. The bondholder has the option of converting the bond into common equity. The bond is used to pay for the equity. Assume a $1,000 par bond is convertible into 50 shares of common equity. Then the investor is paying $20 per share. The convertible bond can be viewed as a bond plus an option to buy 50 shares at $20 per share.

Corporate Bonds

Having discussed some of the factors that affect bond prices, it is useful to examine corporate bonds in more detail.

Corporate Bond Spreads Corporate bonds have a higher promised interest rate than government bonds. This difference in interest rates is called the spread. Table 21.12 shows the average spread between the spot rate on corporate bonds and the spot rate on government bonds for various maturities and ratings. For example, the four-year spot rate on AA corporate bonds was 7.38% and for four-year governments was 6.925%, resulting in a spread of

$$\text{Four-year AA Spread} = 7.38 - 6.925 = 0.455$$

Note that the empirical spread increases with maturity and with a decrease in rating.

Three factors affect the spread:

1. Expected default loss: Some corporate bonds will default, and investors require a higher promised payment to compensate for the expected loss from defaults.
2. Tax premium: Interest payments on corporate bonds are taxed at the state level, whereas interest payments on government bonds are not.

Table 21.12 Corporate Bond Spreads for Industrial Bonds and Various Ratings, 1987–1996

		Spreads		
Maturity	Treasuries	AA	A	BBB
2	6.414	0.414	0.621	1.167
3	6.689	0.419	0.680	1.205
4	6.925	0.455	0.715	1.210
5	7.108	0.493	0.738	1.205
6	7.246	0.526	0.753	1.199
7	7.351	0.552	0.764	1.193
8	7.432	0.573	0.773	1.188
9	7.496	0.589	0.779	1.184
10	7.548	0.603	0.785	1.180

Source: Elton et al. (2001).

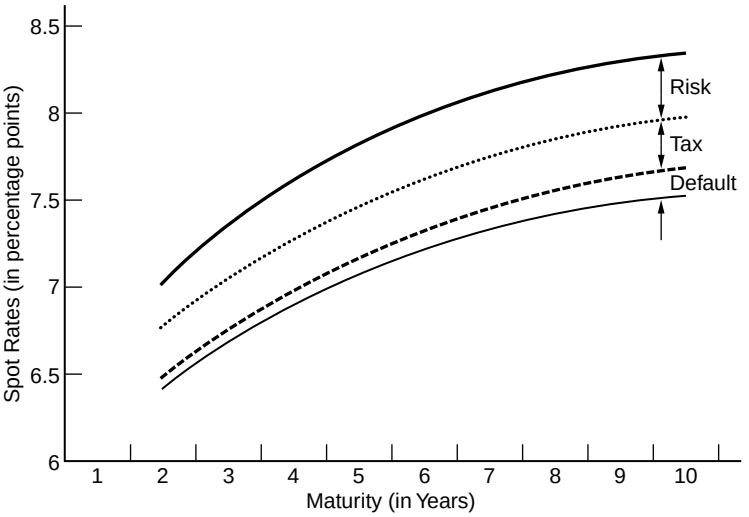


Figure 21.7 Spot rates for A-rated industrial bonds and for Treasuries.

3. Risk premium: The return on corporate bonds is riskier than the return on government bonds, and investors should require a premium for the higher risk. As we will discuss, this occurs because a large part of the risk on corporate bonds is systematic rather than diversifiable.

The first two factors have already been discussed, while the third requires some elaboration. In Chapters 13–16 we presented evidence that an asset with systematic risk requires a higher expected return. Corporate bonds are systematically related to the same factors as common stock. If common stock requires a risk premium, then so should corporate bonds. Furthermore, as shown in Elton, Gruber, Agrawal, and Mann (2001), the sensitivity to the common stock factors increases as rating decreases.

How much of the spread can be attributed to each of the three factors? Figure 21.7 shows the corporate bond spread for A-rated bonds. Most people focus on default as the major determinant of corporate bond spread. For A-rated bonds, relatively little of the spread is explained by the fact that some bonds rated A ultimately default. The fact that corporate bonds are subject to state taxes and government bonds are not explains more of the premium. Finally, the sensitivity of corporate bonds to systematic risk factors and the need to receive a higher return to be compensated for this systematic risk explain the largest part of the spread for A-rated bonds.

Floating Rate Bonds A *floater* is a bond with coupon payments that varies as a function of some interest rate. Consider a floating rate note with a maturity of two years that pays the 6-month Treasury bill rate so that the next coupon is a known amount, and the rate is always fixed at the beginning of each period. Assume the coupon is paid every six months. Thus the coupon that is paid in six months is the current six-month Treasury bill rate. The rate that is paid in one year is the six-month Treasury bill rate that exists six months from now. The rate that is paid in 18 months is the 6-month Treasury bill rate that exists in one year. Assume the bond is riskless, and consider the value of the floating rate

bond in 18 months. In 18 months, there will be one remaining payment of the coupon, and it will be paid at maturity in 24 months, and because it is set in 18 months, it will be known at that time with certainty. Because the bond is riskless by assumption, the value in 18 months is the payment to be made in 24 months brought back at the riskless rate. Because the coupon was fixed at the six-month rate at that time and the discount rate is the six-month rate, the discounted value will be the bond's par value. Let us consider an example. Assume that in 18 months the six-month rate is 6% per year, or 3% per six-month period. If the bond has a par value of $100, the final payment is $103. The value of this final payment in 18 months will be 103/1.03 or $100. The same logic applies at earlier periods. Assume in one year, six-month Treasury bill rates are 5% per year or 2.5% per six-month period. Then, the investor will receive in 18 months a coupon payment worth $2.50 and will have a bond worth $100. The present value as of one year is 102.5/1.025 = $100. Continuing to work back shows that at the time the coupon is reset a riskless floating rate bond paying the Treasury bill rate will always sell at par.

Most floating rate bonds are not riskless. However, their coupon is set at the spread they normally sell above Treasuries. For example, the coupon might be set at the six-month Treasury bill rate plus 2%. As long as the spread remains at a constant 2% per period, the principles discussed earlier hold; namely, at the time the coupon is reset, the bond sells at par.

What is the duration of a floating rate investment? After the reset date the next coupon payment is fixed. Since at the reset date the bond will sell at par, the bond will respond to interest rate changes like a bond that matures at the next reset date. Because between reset dates the bond has the cash flow pattern of a zero or pure discount bond with a maturity equal to the time to reset, and because the duration for zero coupon bonds is the maturity, the duration of a floating rate instrument ignoring any change in spread is the time to reset.

COLLATERAL MORTGAGE OBLIGATIONS

Collateral mortgage obligations (CMOs) are bonds where the underlying asset is a pool of mortgages or a mortgage-backed security. Recall that a mortgage-backed security is a bond where the assets that back it are home mortgages. For example, assume there are 1,000 mortgages, each for $100,000. These mortgages could be pooled and a $100 million bond issued to purchase the mortgages. The payments on the mortgages are then used to pay interest and principal on the bond, and because payments on mortgages are monthly, so are payments on the bond.

The traditional CMO took a mortgage-backed bond and split the income stream into parts and issued bonds against each part. Consider an example. Assume a $100 million mortgage-backed security was the collateral and the bond income stream was split into three parts designated as Class A, Class B, and Class C. Furthermore, assume that there are 100 CMO bonds each with a face value of 1 million and 40 of the 100 bonds are designated as Class A, 30 as Class B, and 30 as Class C. The interest rate on each bond class is the same. However, bonds in Class A receive all of the principal payments, normal and prepayments, until the principal is paid off; then bonds in Class B receive all principal payments until they are paid off, and finally, Class C receives the principal payments. As bonds in a particular class are being paid off, their face value is being reduced so they receive less dollar interest per bond, although the interest rate remains constant. This creates three classes of bonds with very different expected maturities and with very different sensitivity to prepayments. Bonds in Class A have the shortest maturity but are most affected by prepayments, and bonds in Class C have the longest maturity and, at least for a number of years, have very little prepayment risk. Why might the CMO be

more valuable than the underlying mortgage-backed security? It can be more valuable if investors have different desires for bonds of different maturities and possibility for sensitivity to prepayment risk. That the three CMOs might be more valuable than the underlying security seemingly violates the law of one price, which would state that buying bonds in Class A, B and C in proportions that replicate the cash flows of the mortgage-backed bond should cost the same as buying the mortgage-backed security directly. Why is there not a profitable arbitrage? If the value of the aggregate of the CMOs is more than the value of the underlying mortgage-backed security (and it will likely be or it does not pay to create the CMO), then the arbitrage is to create more CMOs. However, the mortgage-backed security used to create the CMO no longer exists.

When subprime mortgages became an important part of the market, a second type of CMO was developed. Subprime mortgages are loans to investors who do not meet normal standards for obtaining mortgages. The CMOs that were backed by subprime mortgages had the following characteristics: first, the pool of subprime mortgages were financed by a number of different classes of bonds. Let us designate the classes as A to J, where the order is from highest rated to lowest rated. Mortgage interest payments went first to Class A, then Class B, and then to Class C, and so on. Assume that Class A is $10 million and the interest rate is 4%, while the lower-rated Class B is $10 million and the interest rate is 4 1/4%. Then the first $400,000 of interest payments goes to Class A, the next $425,000 goes to Class B, and so on.

Defaults go in the opposite direction, with Class J absorbing defaults first and Class I second. Mortgage-backed bonds that are the asset underlying the conventional CMOs were generally guaranteed against default by the issuing organization. However, there was no such guarantee for the new CMOs backed directly by subprime mortgages. One expects some defaults from subprime mortgages, thus there were two forms of credit enhancement. First, unlike conventional CMOs, the interest on the subprime mortgages was more than the interest on the CMO bonds. This was possible because most of the bonds in the CMO were highly rated by the rating agencies, where the subprime mortgages were not, and thus subprime mortgages paid a higher interest rate than the CMOs. Thus the first protection was greater cash flows than were needed to pay the interest on the CMOs. Second, there was an initial pool of cash set aside that could be used to pay interest if the cash flows from the subprimes were insufficient to pay interest on the CMOs. Excess flows from the subprimes went into the safety pool until it reached a certain level, and then it was used to pay off bonds.

THE FINANCIAL CRISIS OF 2008

The financial crisis of 2008 was triggered by massive defaults in the subprime mortgage market, which led to a severe decline in housing prices and a substantial increase in the default of conventional mortgages. This was transmitted to the banking sector and from there to the real sector. To understand this, we first examine how subprime loans worked.

Subprime Loans

There has always been some lending in the subprime market. Prior to 2000 this was a very small fraction of the market, often associated with religious or other charitable groups. The explosion occurred because of a financial innovation, namely, a change in how loans were structured.

The subprime is lending to riskier borrowers. Normally, when lending to riskier borrowers, you compensate for the risk by increasing the interest rate. For this market this

will not work, since higher interest will likely cause borrowers to default. The innovation was to get a greater return by forcing refinancing and capturing a large part of the refinancing costs for the lender. Loans were primarily 2–28's or 3–27's. This means there is a fixed rate for two or three years and then a different rate for the last 27 or 28 years. The interest rate in the first period was relatively low (although in 2007 it averaged 8% on subprime mortgages), but the interest rate in the second period was very high (in 2007 it averaged 13%). This forced homeowners to refinance and gave the lending organization a large additional return because much of the refinancing costs were profit. How was the homeowner to pay for this? As long as house prices continued to rise, this was no problem. The lending organization gave a larger loan, and the homeowner used this extra money to pay the refinancing costs.

Perhaps an example will further clarify how the market worked. Assume an investor borrowed $100,000 on a 3–27 loan, the initial rate was 8%, house prices increased 5% per year, and the initial value of the house was $100,000.[10] Then, in three years, the house would be worth $100,000 $\times$ (1.05) or $115.762. If the refinancing costs were $3,500 with $3,000 pure profit, the lender earns approximately 9%. The homeowner gets a new loan for $103,500, use the $3,500 to pay the refinancing, has $12,262 more in equity, and the lender has a loan with a greater loan-to-value ratio.

The functioning of the market depended on house prices continuing to rise at a rapid rate. When house prices stopped increasing at a rapid rate, there were large defaults in the subprime market because borrowers could not pay for the refinancing and could not afford to pay the higher interest rate. Subprime loans represented about 25% of the market. Large foreclosures in this part of the market transferred to the housing market as a whole and to a general decline in home prices.

The crash in housing prices was exacerbated by what most observers felt was a housing bubble. Housing prices from 1975 to June 2000 had increased by about 1.49% per year. From 2000 to June 2006 they went up 7.7% per year. This was unprecedented. When foreclosures increased rapidly, observers felt the bubble was over, and prices declined rapidly.

Transmittal to the Banks

Major problems in the housing market were transmitted to the banks. There were several reasons for this. First, the Glass-Siegle Act was repealed, allowing commercial banks to engage in underwriting of bonds. Creating subprime CMOs was an extremely profitable business. The subprime CMOs were very large, often in the billions. Banks would accumulate the subprime loans and then create CMOs to back them. However, while they were accumulating subprime mortgages, the mortgages were assets they held, and when they defaulted, it was a loss to the bank. Second, banks were allowed to move assets off their books and finance them with very short-term debt and only a tiny amount of equity, much less equity than for assets on their books. This meant that banks became highly levered (lots of debt relative to equity) and much more levered than regulation would allow if all assets were on the books.

Furthermore, the off-balance-sheet assets were financed by very short-term debt, and if this debt could not be reissued, all of the off-balance-sheet financing went on the balance sheet. Because the equity part of the off-balance-sheet investments was much less than what would be required if it were held directly by the bank, having the off-balance-sheet assets moved to the bank meant the bank might have inadequate capital.

[10]For simplicity I am ignoring any pay down in principal.

Third, the amount of off-balance-sheet assets they held was very difficult to determine, as was how much subprime debt any bank held.

There is a great deal of short-term lending that takes place between banks, and this is needed for credit markets to function. When subprime defaults spiked and house prices fell, banks could not borrow to finance the off-balance-sheet assets because of general pessimism in the markets. Observers knew some banks were in serious trouble and likely to go bankrupt, but not which ones. Thus banks quit lending to one another. This caused chaos in all bond markets.

Furthermore, banks severely contracted all lending because they were so highly levered and were unsure of the size of their losses and felt a need to conserve capital.

Corporations need short-term capital to finance items like inventory and accounts receivable and longer-term capital for machines and plant and equipment. This was not available. Simultaneously, individuals who had suffered large losses in wealth due to decline in home values and great uncertainty cut back on spending. The combination of falling demand and lack of credit led to large layoffs and the slump of 2008.

Credit Default Swaps

Credit default swaps are misnamed. They should be called *default insurance*. A credit default swap insures the purchaser that in the event of a default he will receive the value of his loss. Like any other insurance, the purchaser makes periodic payments to the issuer (say, every quarter). If the bond defaults, the issuer either takes the bond and pays the purchaser the face value of the bond or pays the purchaser the difference between the face value of the bond and the current market value. There are a number of traded instruments that are closely related. These instruments pay off if some percentage of bonds default in a pre-specified pool. These instruments allow the purchaser to hedge against default risk without purchasing credit default insurance.

CONCLUSION

In this chapter, we have introduced bond terminology and the major features of bonds. The only principal feature we introduced but did not devote a section to was the effect of differing coupons. We did not devote a separate section to it because the effect of coupon payments has already been discussed in the tax and maturity sections. In the next chapter, we integrate bond management into portfolio theory.

APPENDIX A

SPECIAL CONSIDERATIONS IN BOND PRICING

The quoted price at which a bond is bought or sold is not the price the customer will pay or receive. The bond selling or purchase price is the quoted price plus accrued interest. Accrued interest is the proportion of interest that has accrued to the bondholder from the last interest payment until the sale or purchase date. For example, assume the quoted sale price is \$96 on a bond paying interest at 10% semiannually. Furthermore, assume there are 181 days between interest payments on the bond and that there are 70 days between the last interest payment and the date the payment will be made for a purchase or sale (settlement date). Then the bond sale price is

$$96 + \frac{70}{181} \times 0.05(100) = \$97.93$$

Rules for calculating accrued interest differ across bond types and countries. The reader calculating the price that will be paid needs to carefully check the rules for the particular bond being purchased.

Note that bond prices and stock prices are quoted on a different basis. Stock prices are quoted at prices at which they are bought or sold. Thus, when a stock pays a dividend, the stock decreases in value by an amount approximately equal to the dividend, and the price drops accordingly. Bonds, on the other hand, are sold at quoted price plus accrued interest. When a bond has an interest payment, the accrued interest becomes zero, but the quoted price remains unchanged.

APPENDIX B

ESTIMATING SPOT RATES

As discussed in the text, spot rates are extremely important in bond valuation and investment decisions, and it is necessary to estimate them. Three techniques have been discussed in the literature. We discuss two of them in this appendix. These two differ in that one of them estimates discrete rates and the other continuous rates.

Consider the following equation relating the price of a bond to the cash flows accruing to the bondholder:

$$P = \frac{c}{(1+S_{01}/2)} + \frac{c}{(1+S_{02}/2)^2} + \frac{c}{(1+S_{03}/2)^3} + \cdots + \frac{1000+c}{(1+S_{0T}/2)^T} \tag{B.1}$$

where

P	is the price of the bond
c	is the coupon
S_{0t}	is the t period spot rate
T	is the number of periods where there are coupon payments
\$1,000	is the principal payment

Alternatively,

$$P = cD_1 + cD_2 + cD_3 + \cdots + (c+1000)D_T \tag{B.2}$$

where

$$D_t = \frac{1}{(1+S_{0t}/2)} \qquad t = 1,\ldots,T$$

The price and the cash flows are known.

As discussed in the text, if we fit Equation (B.2) to multiple bonds simultaneously and recognize that the equation cannot hold exactly for each bond, then Equation (B.2) has the form of a multiple linear regression. D_1 through D_T are the regression coefficients to be estimated. To prevent estimates of forward rates being negative, it is normal to constrain the regression so that the Ds are nonincreasing. Thus D_t is forced to be less than D_{t-1}. The spot rates estimated by this procedure are discrete. Because most bonds pay interest on a semiannual basis, these are spot

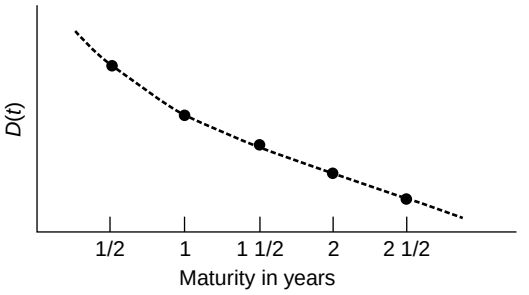


Figure 21.8 Discrete versus continuous discount functions.

rates for cash flows six months apart. Thus $S_{01}/2$ is the spot rate for the first six months, $S_{02}/2$ is for the first 12 months, $S_{03}/2$ is for the first 18 months, and so on. Furthermore, they are rates between specific dates. For example, the six-month intervals could be January to June and July to December. The difficulty with this procedure is that a large number of bonds pay interest on different dates, and the spot rates must be interpolated in some way for use on these dates. Furthermore, bonds with different payment dates cannot be used in the estimation, and thus a fair amount of data are discarded. Carleton and Cooper (1976) suggest the procedure just discussed.

The alternative is to estimate a continuous discount function. The procedure just described estimates $D(t) = 1/(1 + S_{0t}/2)^t$, where the t has integer values such as 1, 2, or 3. $D(t)$ is called a discount function. Consider Figure 21.8, where discount rates for different six-month intervals are plotted. Using the technique just discussed, all we obtain are the points shown. As an alternative, the dashed line could be estimated. This would allow an estimation of discount functions for all maturities. Several forms of equations could be used to approximate the relationship between $D(t)$ and maturity. Because we have assumed some curvature to the relationship, let us approximate it by a quadratic equation. We can write it as

$$D(t) = a_0 + a_1 t + a_2 t^2 \tag{B.3}$$

Once a_0, a_1, and a_2 are known, the spot rates for any time period are known. If the discount function for cash flow in $3\frac{1}{2}$ months is required, then t is set as 3.5/12. The task is thus to estimate a_0, a_1, and a_2. The price of a bond is the present value of its coupon and principal payment. This was written in Equation (B.2) and can be written in compact form as

$$P = \sum_{t=1}^{T} c(t)D(t) \tag{B.4}$$

where $c(t)$ is the coupon payment for all periods before the horizon and the coupon plus principal at the horizon.

Substituting Equation (B.3) into Equation (B.4) yields

$$P = \sum_{t=1}^{T} c(t)\left(a_0 + a_1 t + a_2 t^2\right)$$

Rearranging,

$$P = a_0\left[\sum_{t=1}^{T} c(t)\right] + a_1\left[\sum_{t=1}^{T} tc(t)\right] + a_2\left[\sum_{t=1}^{T} t^2 c(t)\right]$$

Once again, this is in the form of a linear regression. The terms in the brackets and the price are known, and the a_t are regression coefficients. This is the procedure suggested by Schaefer (1981) and McCulloch (1975). The equation used as a discount function by these authors is not exactly equal to that presented in Equation (B.3), but the general procedure is the same.[11]

APPENDIX C

CALCULATING BOND EQUIVALENT YIELD AND EFFECTIVE ANNUAL YIELD

	Normal Frequency of Interest	Quoted Yield	Bond Equivalent Yield	Effective Annual Yield
1. Eurobond	Yearly	y	$2[(1 + y)^{1/2} - 1]$	y
2. Government	Semiannual	y	y	$\left(1 + \frac{y}{2}\right)^2 - 1$
3. Corporate	Semiannual	y	y	$\left(1 + \frac{y}{2}\right)^2 - 1$
4. Ginnie Mae	Monthly	y	$2\left[\left(1 + \frac{y}{12}\right)^6 - 1\right]$	$\left(1 + \frac{y}{12}\right)^{12} - 1$
5. T-bills[a]	None	b	$2[(1 + r)^{365/2N} - 1]$	$(1 + r)^{365/N} - 1$

Variable [a] b = banker's discount yield and $r = b\frac{N}{360}\frac{P_1}{P_0}$.

QUESTIONS AND PROBLEMS

1. Given the following, does the law of one price hold? If not, what action should an investor take?

	Cash Flows in Period		
Bond	1	2	Price
A	100	1,100	970
B	80	1,080	936
C	90	1,090	980

2. Assume a bond with cash flows of $100 each year and a principal payment of $1,000 in five years and a current price of $960. What is

A. Its current yield?

B. Its yield to maturity?

[11]Researchers have used a generalized polynomial curve fitting to estimate this relationship; we select a simple polynomial curve, the quadratic, to illustrate the procedure.

3. Given the following bonds and prices of bonds, what are the spot rates and forward rates?

Bond	Price	1	2	3	4
A	960	1,000			
B	920		1,000		
C	885			1,000	
D	855				1,000

4. Given the cash flows shown below, does the law of one price hold? If not, what is the price of bond *C* that will make it hold?

	Cash Flows in Period		
Bond	1	2	Price
A	80	1,080	982
B	1,100		880
C	120	1,120	1,010

5. Assume the data shown below. What tax rate would make the law of one price hold? Assume that the capital gains tax is one-half the ordinary income tax. Assume that the periods shown are annual and that any capital gain or loss is realized at the time the bond matures.

Bond	1	2	Price
A	80	1,080	985
B	1,100		900
C	120	1,120	1,040

BIBLIOGRAPHY

1. Chang, Ahn, and Thomson, Howard E. "Jump-Diffusion and the Term Structure of Interest Rates," *Journal of Finance*, **43,** No. 1 (March 1988), pp. 155–174.
2. Alexander, Gordon J. "Applying the Market Model to Long-Term Corporate Bonds," *Journal of Financial and Quantitative Analysis*, **XV,** No. 5 (Dec. 1980), pp. 1063–1080.
3. Altman, Edward I. *Investing in Junk Bonds: Inside the High Yield Debt Market* (New York: John Wiley & Sons, 1987).
4. ——. *Default Risk, Mortality Rates and the Performance of Corporate Bonds: 1970–1988* (Charlottesville, VA: Foundation for Research of the Institute for Chartered Financial Analysts, 1989).
5. Altman, Edward, and Keuhne, Brenda. "Default and Returns in the High-Yield Bond Market: 2011 in Review," NYU Salomon Center (Feb. 3, 2012).
6. Ananthanarayanan, A. L., and Schwartz, Eduardo S. "Retractable and Extendible Bonds: The Canadian Experience," *Journal of Finance*, **35,** No. 1 (March 1980), pp. 31–48.
7. Asquith, Paul, Mullins, David, and Wolff, Eric. "Original Issue High Yield Bonds: Aging Analysis of Defaults, Exchanges and Calls," *Journal of Finance*, **44** (Sept. 1989), pp. 923–953.
8. Atkinson, T. R. *Trends in Corporate Bond Quality* (New York: National Bureau of Economic Research, 1967).

9. Balduzzi, Pierluigi, Elton, Edwin J., and Green, T. Clifton. "Economic News and the Yield Curve: Evidence from the US Treasury Market," *Journal of Financial and Quantitative Analysis*, **36** (Dec. 2001) pp. 523–543.
10. Black, Fischer, and Cox, John C. "Valuing Corporate Securities: Some Effects of Bond Indenture Provisions," *Journal of Finance*, **31** (May 1976), pp. 351–367.
11. Black, Fischer, and Scholes, Myron. "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, **81** (May/June 1973), pp. 637–654.
12. Blume, Marshall E., and Keim, Donald B. "Risk and Return Characteristics of Lower Grade Bonds," *Financial Analysts Journal* (July/Aug. 1987), pp. 26–33.
13. Brennan, M. J., and Schwartz, E. S. "Convertible Bonds: Valuation and Optimal Strategies for Call and Conversion," *Journal of Finance*, **32** (Dec. 1977), pp. 1699–1715.
14. ——. "Conditional Predictions of Bond Prices and Returns," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 405–416.
15. ——. "An Equilibrium Model of Bond Pricing and a Test of Market Efficiency," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 3 (Sept. 1982), pp. 301–330.
16. ——. "Bond Pricing and Market Efficiency," *Financial Analyst Journal*, **38,** No. 5 (Sept./Oct. 1982), pp. 49–56.
17. Brown, Stephen J., and Dybvig, Philip H. "The Empirical Implications of the Cox, Ingersoll, Ross Theory of the Term Structure of Interest Rates," *Journal of Finance*, **41,** No. 3 (July 1986), pp. 617–630.
18. Buse, A. "Interest Rates, the Meiselman Model and Random Numbers," *Journal of Political Economy*, **LXXV** (Feb. 1967), pp. 49–62.
19. Cagan, Philip. *Changes in the Cyclical Behavior of Interest Rates* (New York: National Bureau of Economic Research, 1966).
20. Campbell, John Y. "A Defense of Traditional Hypotheses about the Term Structure of Interest Rates," *Journal of Finance*, **41,** No. 1 (March 1986), pp. 183–193.
21. Carleton, W. T., and Cooper, I. A. "Estimation and Uses of the Term Structure of Interest Rates," *Journal of Finance*, **31** (Sept. 1976), pp. 1067–1083.
22. Chambers, Donald R., Carleton, Willard T., and Waldman, Donald W. "A New Approach to Estimation of the Term Structure of Interest Rates," *Journal of Financial and Quantitative Analysis*, **19,** No. 3 (Sept. 1984), pp. 233–252.
23. Chance, Don M. "Default Risk and the Duration of Zero Coupon Bonds," *Journal of Finance*, **45,** No. 1 (March 1990), pp. 265–274.
24. Conard, Joseph W. *Introduction to the Theory of Interest* (Berkeley: University of California Press, 1959).
25. Constantinides, George M., and Ingersoll, Jonathan E., Jr. "Tax Effects and Bond Prices," *Journal of Finance*, **37,** No. 2 (May 1982), pp. 349–351.
26. Cox, John C., Ingersoll, Jonathan E., and Ross, Stephen. "An Analysis of Variable Rate Loan Contracts," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 389–404.
27. Culbertson, John M. "The Term Structure of Interest Rates," *Quarterly Journal of Economics*, **LXXI** (Nov. 1957), pp. 485–517.
28. Dermoday, Jaime Cuevas, and Prisman, Eliezen Zeev. "Term Structure Multiplicity and Clientele in Markets with Transactions Costs and Taxes," *Journal of Finance*, **43,** No. 4 (Sept. 1989), pp. 893–911.
29. Duentz, Mark L., and Mahoney, James M. "Using Duration and Convexity in the Analysis of Callable Bonds," *Financial Analyst Journal*, **44,** No. 3 (May/June 1988), pp. 53–72.
30. Elton, Edwin J., and Green, T. Clifton. "Tax and Liquidity in Pricing of Government Bonds," *Journal of Finance*, **53,** No. 5 (Oct. 1998), pp. 1533–1562.
31. Elton, Edwin J., Gruber, Martin J., Agrawal, Deepak, and Mann, Christopher. "Explaining the Rate Spread on Corporate Bonds?" *Journal of Finance*, **LVI,** No. 1 (Feb. 2001), pp. 247–279.
32. Fama, E. "Forward Rates as Predictors of Future Spot Rates," *Journal of Financial Economics*, **3** (Oct. 1976), pp. 361–377.
33. ——. "The Information in the Term Structure," *Journal of Financial Economics*, **13,** No. 4 (Dec. 1984), pp. 509–528.

34. ——. "Term Premiums in Bond Returns," *Journal of Financial Economics*, **13,** No. 4 (Dec. 1984), pp. 529–546.
35. Fisher, Lawrence. "Determinants of Risk Premiums on Corporate Bonds," *Journal of Political Economy*, **67** (June 1959), pp. 217–237.
36. Fitzpatrick, J. D., and Severiens, J. T. "Hickman Revisited: The Case for Junk Bonds," *Journal of Portfolio Management*, **4,** No. 4 (Summer 1978), pp. 53–57.
37. Fraine, H. G., and Mills, R. H. "The Effect of Defaults and Credit Deterioration on Yields of Corporate Bonds," *Journal of Finance*, **16** (Sept. 1961), pp. 423–433.
38. Galai, Dan. "Pricing of Optionable Bonds," *Journal of Business Finance*, **7,** No. 3, (Sept. 1983), pp. 323–337.
39. Hansen, L. P., and Hodrick, R. J. "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis," *Journal of Political Economy*, **88** (Oct. 1980), pp. 829–853.
40. Hickman, W. Braddock. *Corporate Bond Quality and Investor Experience* (Princeton, NJ: Princeton University Press and the National Bureau of Economic Research, 1958).
41. Hill, J. H., and Post, L. A. *The 1977–78 Lower-Rated Debt Market: Selectivity, High Yields, Opportunity* (New York: Smith Barney Harris Upham, 1978).
42. Ho, Thomas, and Singer, Ronald F. "The Value of Corporate Debt with a Sinking-Fund Provision," *Journal of Business*, **57,** No. 3 (Oct. 1984), pp. 315–336.
43. Johnson, Ramon, "Term Structure of Corporate Bond Yields as a Function of Risk of Default," *Journal of Finance*, **22** (May 1967), pp. 313–345.
44. Kane, Alex, and Marcus, Alan F. "Valuation and Optimal Exercise of the Wild Card Option in the Treasury Bond Futures Market," *Journal of Finance*, **41,** No. 1 (March 1986), pp. 195–207.
45. Kessel, Reuben H. *The Cyclical Behavior of the Term Structure of Interest Rates* (New York: National Bureau of Economic Research, 1965).
46. Lang, Richard, and Rasche, Robert. "Debt-Management Policy and the Own Price Elasticity of Demand for U.S. Government Notes and Bonds," *Federal Reserve Bank of St. Louis Review*, **59** (Sept. 1977), pp. 8–22.
47. Lee, Wayne, Maness, Terry, and Tuttle, Donald. "Nonspeculative Behavior and the Term Structure," *Journal of Financial and Quantitative Analysis*, **15** (March 1980), pp. 53–83.
48. Litzenberger, Robert H., and Rolfo, Jacques. "An International Study of Tax Effects on Government Bonds," *Journal of Finance*, **39,** No. 1 (March 1984), pp. 1–22.
49. ——. "Arbitrage Pricing, Transaction Costs and Taxation of Capital Gains: A Study of Government Bonds with the Same Maturity Date," *Journal of Financial Economics*, **13,** No. 3 (Sept. 1984), pp. 337–352.
50. Lutz, Friedrich A. "The Structure of Interest Rates," *Quarterly Journal of Economics*, **LV** (Nov. 1940), pp. 36–63.
51. Marsh, Terry. "Equilibrium Term Structure Models: Test Methodology," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 421–434.
52. McConnell, John J., and Schwartz, Eduardo S. "LYON Taming," *Journal of Finance*, **41,** No. 3 (July 1986), pp. 561–576.
53. McCulloch, J. H. "Measuring the Term Structure of Interest Rates," *Journal of Business*, **44** (Jan. 1971), pp. 19–31.
54. ——. "An Estimate of the Liquidity Premium," *Journal of Political Economy*, **83** (Feb. 1975), pp. 95–119.
55. ——. "The Tax-Adjusted Yield Curve," *Journal of Finance*, **30** (June 1975), pp. 811–830.
56. Malkiel, Burton G. "Expectations, Bond Prices, and the Term Structure of Interest Rates," *Quarterly Journal of Economics*, **LXXVI** (May 1962), pp. 197–218.
57. ——. *The Term Structure of Interest Rates* (Princeton, NJ: Princeton University Press, 1966).
58. Meiselman, David. *The Term Structure of Interest Rates* (Englewood Cliffs, NJ: Prentice Hall, 1962).
59. Merton, Robert. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *Journal of Finance*, **29** (May 1974), pp. 449–470.
60. Modigliani, Franco, and Sutch, Richard. "Innovations in Interest Rate Policy," *American Economic Review*, **LVI** (May 1966), pp. 178–197.

61. ——. "Debt Management and the Term Structure of Interest Rates: An Empirical Analysis of Recent Experience," *Journal of Political Economy*, **75** (Supplement, Aug. 1967), pp. 569–589.
62. Pinches, G. E., and Mingo, K. A. "A Multivariate Analysis of Industrial Bond Ratings," *Journal of Finance*, **28** (March 1973), pp. 1–32.
63. Piros, Christopher D. "Taxable vs. Tax-Exempt Bonds: A Note on the Effect of Uncertain Taxable Income," *Journal of Finance*, **42,** No. 2 (June 1987), pp. 447–451.
64. Rao, Ramesh K. S. "The Impact of Yield Changes on the Systematic Risk of Bonds," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1982), pp. 115–128.
65. Roll, Richard. *The Behavior of Interest Rates* (New York: Basic Books, 1970).
66. ——. "After-Tax Investment Results from Long-Term vs. Short-Term Discount Coupon Bonds," *Financial Analysts Journal*, **40,** No. 1 (Jan./Feb. 1984), pp. 43–54.
67. Sarig, Oded, and Warga, Arthur. "Some Empirical Estimates of the Risk Structure of Interest Rates," *Journal of Finance*, **44,** No. 5 (Dec. 1989), pp. 1351–1360.
68. Schaefer, S. M. "Measuring a Tax Specific Term Structure of Interest Rates in the Market for British Government Securities," *Economic Journal*, **91** (June 1981), pp. 415–438.
69. ——. "Tax-Induced Clientele Effect in the Market for British Government Securities: Placing Bounds on Security Values in an Incomplete Market," *Journal of Financial Economics*, **X,** No. 2 (July 1982), pp. 121–160.
70. Shea, Gary S. "Pitfalls in Smoothing Interest Rate Term Structure Data: Equilibrium Models and Spline Approximations," *Journal of Financial and Quantitative Analysis*, **19,** No. 3 (Sept. 1984), pp. 253–270.
71. Smith, Clifford, and Warner, Jerold. "On Financial Contracting: An Analysis of Bond Covenants," *Journal of Financial Economics*, **7** (June 1979), pp. 115–161.
72. Sundaresan, M. "Constant Absolute Risk Aversion Preferences and Constant Equilibrium Interest Rates," *Journal of Finance*, **39,** No. 1 (March 1983), pp. 205–212.
73. ——. "Consumption and Equilibrium Interest Rates in Stocastic Production Economies," *Journal of Finance*, **39,** No. 1 (March 1984), pp. 77–92.
74. Telser, L. G. "A Critique of Some Recent Empirical Research on the Explanation of the Term Structure of Interest Rates," *Journal of Political Economy*, **75** (Supplement, Aug. 1967), pp. 546–561.
75. Torous, Walter N. "Differential Taxation and the Equilibrium Structure of Interest Rates," *Journal of Business Finance*, **9,** No. 3 (Sept. 1985), pp. 363–385.
76. Vanderhoof, Irwin T. F., Tenenbein, Albert A., and Verni, R. "The Risk of Asset Default," *Report of the Society of Actuaries*, C1 Task Force of the Committee on Valuation and Related Areas (1989).
77. Van Horne, James. "Interest-Rate Risk and the Term Structure of Interest Rates," *Journal of Political Economy*, **LXXIII** (Aug. 1965), pp. 344–351.
78. ——. "Interest-Rate Expectations, the Shape of the Yield Curve, and Monetary Policy," *Review of Economics and Statistics*, **XLVIII** (May 1966), pp. 211–215.
79. ——. "The Expectations Hypothesis, the Yield Curve, and Monetary Policy: Comment," *Quarterly Journal of Economics*, **LXXIX** (Nov. 1965), pp. 664–668.
80. Van Horne, James C. "Implied Tax Rates and the Valuation of Discount Bonds," *Journal of Business Finance*, **6,** No. 2 (June 1982), pp. 145–159.
81. Van Horne, James C., and Bowers, David A. "The Liquidity Impact of Debt Management," *Southern Economic Journal*, **XXIV** (April 1968), pp. 526–537.
82. Walsh, Carl E. "A Rational Expectations Model of Term Premium with Some Implications for Empirical Asset Demand Equations," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 63–83.
83. Wood, John H. "Expectations, Error, and the Term Structure of Interest Rates," *Journal of Political Economy*, **LXXI** (April 1963), pp. 160–171.
84. Zwick, Burton. "Yield on Privately Placed Corporate Bonds," *Journal of Finance*, **35,** No. 1 (March 1980), pp. 23–30.

22

The Management of Bond Portfolios

In the previous chapter, we discussed the determination of interest rates and the characteristics of bonds that affect their return and value. In this chapter we discuss bond portfolio management. Modern portfolio theory has made less of an impact on bond management than it has on common equity management. Furthermore, some of the portfolio management techniques used in bond management are specific to the bond area and not outgrowths of modern portfolio theory. In this chapter we discuss the techniques specifically developed for the bond area as well as applications of general portfolio theory to the bond area.

The chapter is divided into four parts. First we discuss the major source of risk facing bond managers, changes in the yield curve, and measures used to examine a bond's sensitivity to this source of risk. Next we discuss ways of constructing a bond portfolio to insulate against this risk. These are normally referred to as passive portfolio strategies, although, as we will see, they generally involve actively adjusting the portfolio. Next we will discuss active bond management. We discuss both techniques developed specifically for active bond management and bond management in a modern portfolio theory context, discussing first estimation of expected return and then estimation of the variance–covariance structure. Finally, we discuss bond and interest rate swaps.

DURATION

The return on a bond has two components: interest income and capital gains or losses caused by a change in price. A price change can come about because of the passage of time or as a result of a shift in the yield curve. In what follows, it will be convenient to assume interest is paid annually. Furthermore, we assume a flat yield curve with all spot rates equal to i. In Appendix A we discuss the minor modification needed for bonds paying semiannual or monthly interest. We also discuss the changes needed when there is an upward-sloping yield curve.

Price Change due to Passage of Time

Consider first a price change due to the passage of time. Assume a flat yield curve with an interest rate of 10%. Now consider a pure discount bond with three years to maturity. The

price of a pure discount three-year bond that pays $1,000 at maturity is

$$P_3 = \frac{1{,}000}{(1.10)^3} = \$751.31$$

assuming that spot rates remain unchanged over the first year.[1] Then, at time 1, this bond must have the same yield as a two-year bond and thus have a price of

$$P_2 = \frac{1{,}000}{(1.10)^2} = \$826.45$$

This price change would occur over the year. The price change over the year is $P_2 - P_3 = \$75.14$, which results in a rate of return of

$$\frac{P_2 - P_3}{P_3} = 10\%$$

The effect of the passage of time on the price of a bond should be easy to understand for pure discount bonds. Because pure discount bonds do not pay any interest, the full return is due to a change in price. Coupon-paying bonds also can have an expected price change due to the passage of time. There are a large number of bonds that are comparable in every way, except that they offer different coupons. These bonds must offer similar returns to investors. Thus, for these bonds, there are anticipated price changes. For example, a 4% coupon bond will sell at a discount and offer an expected price increase if current interest rates for a similar bond are 10%. Most bonds include an anticipated price change as part of their return.

Unanticipated Price Change

The other cause of a price change is a change in future expectations concerning interest rates (an unanticipated shift in the yield curve). Assume that the yield curve shifts and the new interest rate for all maturities is 14%. Further assume that the shift takes place immediately. In this case the three-year, pure discount bond would have a new price of

$$P_3' = \frac{\$1{,}000}{(1.14)^3} = \$674.97$$

This results in a price change of

$$P_3' - P_3 = -\$76.34$$

If the yield curve remains constant over time or if expectations remain constant, the price change due to the passage of time is easy to calculate. The price change due to an unanticipated change in the yield curve is different.

If we knew how expectations concerning future interest rates would shift over time and others did not, then we would be able to calculate the price change of each bond and put all of our money into the bond with the highest total return. However, this is not possible; the best that we can do is to calculate the sensitivity of each bond to a shift in the yield curve.

[1]This is the simplest example that can be constructed. It assumes a flat yield curve. The principle being demonstrated also holds under more complex shapes of the yield curve.

Sensitivity to Shifts in the Yield Curve

In earlier chapters we calculated a measure called beta to measure a common equity security's sensitivity to changes in an index. An analogous measure is calculated for bonds: it is called *duration*. Duration is a measure of the sensitivity of the price of a bond to a change in interest rates. More specifically, minus duration times the proportional change in 1 plus the interest rate is equal to the unanticipated return due to a change in price.[2]

In symbols,

$$R_u = -D\,\Delta i \tag{22.1}$$

where

- i is the interest rate
- R_u is the unanticipated return due to a change in the interest rate
- D is duration
- Δi is the proportional change in 1 plus the interest rate $\left(\dfrac{d(1+i)}{1+i}\right)$

Note that we have dropped subscripts on the interest rate and have been referring to "the interest rate" as if there is a single rate that does not depend on maturity. Furthermore, to emphasize this change, we use the symbol *i*. This is in contrast to earlier sections, where we were clearly specifying the time horizon of the interest rate. For simplicity, we are assuming a single rate for all maturities. A single rate is an assumption of a flat yield curve. In the appendix, we generalize the analysis.

To understand duration, consider a pure discount bond that matures in T years. Coupon-paying bonds can be considered as combinations of pure discount bonds. Thus understanding duration for pure discount bonds will help us understand it for coupon bonds. Let P_0 be the current price of a pure discount bond that pays \$1,000 in T years. If i is the yearly interest rate, then

$$P_0 = \frac{\$1{,}000}{(1+i)^T} \tag{22.2}$$

We derive the duration for this bond in the following section. The reader uninterested in the derivation can skip to the end of the dotted section.

Equation (22.2) can be written as

$$P_0 = (1{,}000)(1+i)^{-T}$$

[2]Security firms generally calculate a slight variation of this formula. Equation (22.1) is

$$R_u = -D\Delta i = -D\frac{d(1+i)}{(1+i)}$$

where $d(1 + i)$ is the change in 1 plus the interest rate. Security firms divide duration by $(1 + i)$ and call this adjusted or modified duration. Thus modified duration (D_A) is

$$D_A = \frac{D}{1+i}$$

and Equation (22.1) becomes

$$R_u = -D_A d(1+i) = -D_A di$$

Recall that the derivative of X^N is NX^{N-1}dX. Thus

$$dP_0 = 1{,}000(-T)(1+i)^{(-T-1)}\,d(1+i)$$

Rearranging yields

$$dP_0 = \frac{-1{,}000T}{(1+i)^T}\,\frac{d(1+i)}{1+i}$$

Note that $1{,}000/(1+i)^T$ is the price of the bond P_0; thus

$$dP_0 = -TP_0\,\frac{d(1+i)}{(1+i)}$$

Dividing both sides by P_0, we have

$$\frac{dP_0}{P_0} = \frac{-TP_0}{P_0}\,\frac{d(1+i)}{(1+i)} = -T\,\frac{d(1+i)}{(1+i)}$$

The change in price divided by price dP_0/P_0 is the return due to an unanticipated change in the interest rate; $d(1+i)/(1+i)$ is the proportional change in 1 plus the interest rate.

Comparing this expression to Equation (22.1) and recognizing that $R_u = \dfrac{dP_0}{P_0}$, we see that $D = T$. Thus the duration on a pure discount bond is its maturity.

For a pure discount bond such as that presented in Equation (22.2), duration is equal to its maturity. Thus, given the assumption of a flat yield curve, the sensitivity of a pure discount bond to a change in the yield curve should be directly proportional to its maturity. When the change in the interest rate divided by 1 plus the interest rate is equal to 1%, the change in the price of a pure discount bond with a maturity of one year should be 1%, and the change in price of a pure discount bond with a maturity of five years should be 5%, and so on.

Table 22.1 illustrates these ideas. The bonds in Table 22.1 are pure discount bonds with the maturity shown in the first column. All bonds are assumed to return a principal of $1,000 at the horizon. The prices are shown in the next two columns under two alternative interest rate assumptions, 10% and 10.11%. The change in interest rate between the two columns is 0.1011 − 0.10 or 0.0011. The percentage change in 1 plus the interest rate is 0.0011/(1.10) or 0.1%. The percentage change in price from the second to the third column should be minus duration times this 0.1 figure. Because for pure discount bonds, duration is maturity, the last column should be minus (0.1) times maturity, and it is. The analysis is derived for very small changes in interest rates, and it holds exactly for a very small change

Table 22.1 The Effect of a Change in Interest Rates on the Price of a Pure Discount Bond

Maturity (year)	Price		Percentage Change in Price
	$i = 10\%$	$i = 10.11\%$	
1	$909.09	$908.18	−0.1
2	$826.45	$824.80	−0.2
3	$751.31	$749.07	−0.3
4	$683.01	$680.29	−0.4
5	$620.92	$617.83	−0.5

in rates. For large changes in rates, the duration measure provides only an approximation of the actual percentage change in prices. However, the approximation is a good one.

Coupon-paying bonds can be viewed as combinations of pure discount bonds. Consider a bond with two payments, one in 5 years and one in 10 years. If we consider each payment separately, and designate the return on the payment in 5 years due to an unanticipated change in interest rates as R_u^5 and the return on the payment in 10 years due to an unanticipated change in interest rates as R_u^{10}, we have

$$R_u^5 = -5\,\Delta i$$
$$R_u^{10} = -10\,\Delta i$$

The bond with two payments can be viewed as a portfolio of the 5-year payment and the 10-year payment. Let P_5 be the present value of the 5-year payment and P_{10} be the present value of the 10-year payment, P_0 be the value of the bond, and R_u be the unanticipated return on the portfolio.

In earlier chapters, we showed that the return on a portfolio is a weighted average of the return on the assets composing that portfolio and that the weights are the fraction of the money invested in the asset. The same principles apply here. Thus the unanticipated return on the portfolio is simply the sum of the fraction of the portfolio invested in each payment times the unanticipated return on the appropriate payment. The fraction invested in each payment is the present value of that payment divided by the price of the portfolio. Thus[3]

$$R_u = \left(\frac{P_5}{P_0}\right)R_u^5 + \left(\frac{P_{10}}{P_0}\right)R_u^{10}$$

Substituting in for R_u^5 and R_u^{10} yields

$$R_u = \frac{P_5}{P_0}(-5\,\Delta i) + \frac{P_{10}}{P_0}(-10\,\Delta i)$$
$$= -\left[\frac{P_5}{P_0}(5) + \frac{P_{10}}{P_0}(10)\right]\Delta i$$

Thus the duration of a bond with two payments is a weighted average of the maturity of each payment, where the weights are the proportion of the current value of the bond attributable to that payment. If the 5- and 10-year payment each contributed equally to the current value of the bond, then the duration would be $7\frac{1}{2}$ years. This can be generalized to T payments. The present value of a payment made in period t is $C(t)/(1 + i)^t$, where $C(t)$ is the payment in period t. If P_0 is the price of the bond, then the fraction of the present value of each payment is $[C(t)/(1 + i)^t]/P_0$. Each weight is multiplied by the duration of the payment that is its maturity. Thus the duration of a T-period bond with payments in each period is

$$D = \frac{\frac{C(1)}{(1+i)}}{P_0}1 + \frac{\frac{C(2)}{(1+i)^2}}{P_0}2 + \cdots + \frac{\frac{C(T)}{(1+i)^T}}{P_0}T$$

$$D = \frac{\sum_{t=1}^{T}\frac{tC(t)}{(1+i)^t}}{P_0} \tag{22.3}$$

[3]For the measure of duration under discussion, duration is additive only if there is a flat yield curve, which is what we have assumed.

Table 22.2 Duration of Bonds with Different Maturities and Coupons[a]

	Years to Maturity		
Coupon	3	5	10
4	2.88	4.57	7.95
6	2.82	4.41	7.42
8	2.78	4.28	7.04
10	2.74	4.17	6.76
12	2.70	4.07	6.54
14	2.66	3.99	6.36

[a]The analysis assumes $i = 10\%$ and annual payment of coupons.

Notice that the duration for coupon-paying bonds is less than the maturity. Up to now we have assumed that the yield curve is flat and that a shift takes place in the flat yield curve. Many other assumptions could be made. Different assumptions change the definition of duration. We have set out several of these in Appendix A at the end of the chapter. Researchers have compared these measures to see which seems to be the most accurate representation of a bond's sensitivity to a change in interest rates. The surprising result is that the one we have presented in this chapter, which was the first one ever derived, seems to do well in explaining unanticipated returns. Its performance and simplicity help explain why this measure is the one most widely used in practice.

Table 22.2 shows the duration on a number of bonds with different maturities and different coupons. Notice how the duration of a bond is much shorter than its maturity, especially for bonds with long maturities.

Equation (22.3) shows that the duration of a bond is affected by the maturity of the bond, its coupon, and the interest rate. Holding changes in other variables constant, we see the following:

1. An increase in the coupon lowers duration. This is illustrated in Table 22.2, and the logic behind it is easy to understand. As the coupon is increased, the value of the earlier cash flows increases relative to the present value of the terminal cash flow. This increases the weight of the early cash flows and lowers duration.
2. An increase in the interest rate lowers duration. The greater the interest rate, the less important are cash flows far in the future relative to near-term flows. The greater the weight on near-term flows, the lower the duration.
3. In general, the longer the maturity, the greater the duration. This is illustrated in Table 22.2.[4]

Although this ends our presentation of the concept of duration, we return to duration and use it as a tool in bond portfolio management in the later section of the chapter.

Convexity

In recent years, there has been an increased realization that although duration works well in explaining changes in price for small shifts in the yield curve, it does not work nearly

[4]A decrease would be rare. Only for deep discount bonds (low coupon) could duration shorten with an increase in maturity. If the coupon is sufficiently low, then receiving the principal later (longer maturity) may lower price by more than price is increased because of the extra coupons. If price is lowered, then the weight on the early payments (which is the present value of the payment divided by price) will be increased, and duration can be shortened.

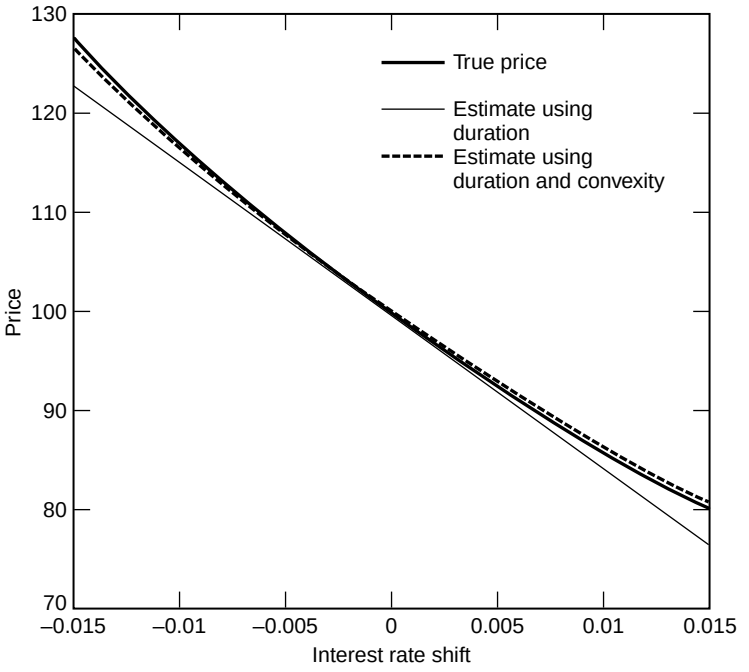


Figure 22.1 Actual price change and estimated price change.

as well for larger shifts. Duration assumes that the percentage price change is proportional to the percentage change in 1 plus the interest rate. This approximation becomes increasingly bad for large changes in interest rates.

A correction term has been developed that is generally known as *convexity*. The term *convexity* arises from the fact that percentage price change approximates a convex function rather than a linear function of changes in 1 plus the interest rate (see Figure 22.1). The derivation of convexity is described in Appendix D, whereas the formula for unexpected return is given here:

$$R_u = -D\,\Delta i + C(\Delta i)^2$$

$$C = \left(\frac{1}{2}\right)\frac{\sum_{t=1}^{T}\frac{t(t+1)C(t)}{(1+i)^t}}{P_0} \tag{22.4}$$

and D is as described in Equation (22.3).

As an illustration of the use of convexity, let us return to the example presented in Table 22.1. Consider the 5-year pure discount bond, which pays \$1,000 at maturity. Change the assumption in the table to a larger change in interest rates; in particular, assume interest rates change from 10% to 12.2%. At a 10% interest rate, the price of the 5-year pure discount bond is

$$\$620.92 = \left(\frac{1{,}000}{(1.10)^5}\right)$$

whereas at a 12.2% interest rate, the price is \$562.39.

The rate of price change as interest rates rise from 10% to 12.2% is

$$\frac{P_{12.2} - P_{10}}{P_{10}} = \frac{562.39 - 620.92}{620.92} = -0.094 \text{ or } -9.4\%$$

The duration on the bond is 5 years, whereas $\Delta i = 0.022/1.10 = 0.02$. If we estimated the unexpected rate of return on the bond just using duration [Equation (22.1)], we would estimate it as

$$R_u = -5(0.02) = -0.10$$

This is a 6.4% error. To obtain a better estimate, we wish to apply Equation (22.4), which corrects the duration measure for convexity.

The convexity on this bond is

$$C = \left(\frac{1}{2}\right)\frac{\frac{5(6)\,1{,}000}{(1.10)^5}}{\$620.92} = 15$$

$$R_u = -5(0.02) + 15(0.02)^2 = -0.10 + 0.006 = -0.094 \text{ or } -9.4\%$$

The convexity measure has produced an exact estimate in this case. In general, even using duration and convexity, the estimate will only be an approximation, though often a very good one.

As a second example, consider Figure 22.1, which plots the actual price of a bond when different flat yield curves are assumed. The bond prices being plotted are for an eight-year bond with a 10% coupon that pays interest semiannually. Also plotted on the curve is the estimated price of the bond using duration alone (the straight line) and using duration plus convexity (the dashed curve). For small changes in the yield curve, the actual price change is closely matched by both, the estimate using duration alone and the estimate using duration plus convexity.[5] For large price changes, the introduction of convexity improves the estimation.

So far we have graphed only the relationship between price and yield for bonds without call features. For these bonds the relationship has the nice curved shape shown in Figure 22.1. The curved shape is known as the convexity, and for bonds without options such as those depicted in Figure 22.1, it is called *positive convexity*. When bonds have option features, the relationship between price and yield is not so simple. Figure 22.2 plots the relationship between price and yield for a callable bond. This relationship has negative convexity for yield below 10% but positive convexity for yields above. The reason for the shape for yield below 10% is easy to understand. As the price of the bond exceeds the call price, it pays the corporation to call. Investors knowing this will not pay much above the call price for the bond for fear that the corporation will call. Thus below the yield of 10%, the price curve flattens out.[6]

---

[5]The plots of price using duration and using duration plus convexity were obtained as follows. When the yield changes, the bond price changes. The unanticipated return is the change in price divided by the preshift price or

$$R_u = \frac{\Delta P}{P}$$

Combining this equation with (22.4), we have

$$\frac{\Delta P}{P} = -D\,\Delta_i + C\,\Delta_i^2 \qquad \text{or} \qquad \Delta P = -DP(\Delta_i) + CP(\Delta_i^2)$$

The plots were obtained with C equal to zero when calculating the approximation using duration alone or its calculated value when calculating the approximation using duration plus convexity.

[6]The corporation may not call exactly when the price exceeds the call price because of a belief that rates will fall even further. Thus it is rational for the bonds to trade slightly above the call price, and they do.

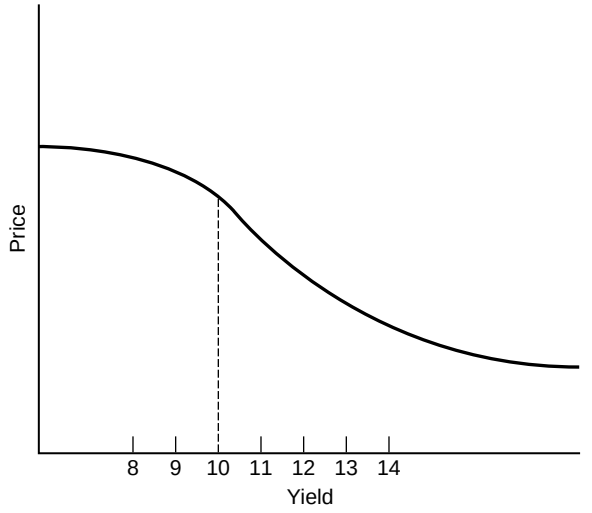


Figure 22.2 The relationship between yield and price for a callable bond.

PROTECTING AGAINST TERM STRUCTURE SHIFTS

Shifts in the term structure are viewed by most managers as the major sources of risk to bond portfolios. Just as shifts in the market systematically affect all equity prices, shifts in the term structure affect all bond prices.

Two techniques have been devised to try to insulate a portfolio from shifts in the term structure. These techniques are known as *exact matching* and *immunization*.

Exact Matching or Dedication

Exact matching involves finding the lowest cost portfolio that produces cash flows exactly matching the outflows that are financed by the investment. Consider the example shown in Table 22.3. In this example we assume it is necessary to meet flows of $100, $1,000, and $2,000 over the next three years. These cash flows might be needed to meet pension payments. The bond portfolio is the investment used to meet these obligations. An exact matching program would determine a bond portfolio of one-, two-, and three-year bonds so that the coupons plus principal exactly match the three flows mentioned.

Portfolio A in Table 22.3 is a portfolio that is cash flow matched. Most investment organizations also consider portfolios with surplus cash flows in early periods that can be used to meet liabilities in latter periods as cash flow matched. This is illustrated by portfolio B in Table 22.3. In this example, $100 of the inflow of $195 in period 1 is used to

Table 22.3 Cash Flow Matched Portfolios

	Period		
	1	2	3
Liability	$100	$1,000	$2,000
Portfolio A	$100	$1,000	$2,000
Portfolio B	$195	$900	$2,000

meet the liability of $100 in period 1, and $95 is invested and carried forward to period 2 to finance the shortfall of $100 in period 2. As long as $5 in interest can be earned on the period 1 surplus, the portfolio is cash flow matched.

In Appendix B, we discuss a procedure for determining a portfolio to accomplish this matching as well as variations on this procedure that can lead to lower-cost portfolios. Exact matching programs are a passive investment program. Once the portfolio is determined, no additional changes are required, even if the yield curve changes in dramatic ways. The performance of the portfolio is insensitive to interest rate shifts in the sense that it meets a fixed set of obligations regardless of changes in the yield curve. In practice, when the yield curve shifts, there may well be profitable bond swaps, and a firm using an exact matching program would use the procedures of Appendix C to evaluate these swaps.

There are two risks with exact matching programs. First, the cash flows may not materialize because of bonds defaulting or being called. Second, if the strategy involves cash carry forward (portfolio B in Table 22.3), then there is risk that return on the funds carried forward will be inadequate. Nevertheless, the manager is reasonably assured of meeting the liabilities even with shifts in the yield curve.

Immunization

The second category of techniques for protecting against interest rate shifts is immunization programs. Earlier we introduced duration as a measure of the sensitivity of a bond or a portfolio of bonds to interest rate shifts. Immunization theory attempts to eliminate sensitivity to shifts in the term structure by matching the duration of the assets to the duration of the liabilities. Thus, if duration is truly a measure of sensitivity to interest rate shifts, a shift in the term structure will have the same impact on the present value of both assets and liabilities and will leave unchanged the ability of the program to meet any obligations. If interest rates rise, the present value of assets and liabilities will fall by the same amount. Similarly, if interest rates fall, then the value of the assets and liabilities will rise by the same amount. Perhaps an analogy to beta is helpful. If a liability had a beta of 1.5, then purchasing an asset with a beta of 1.5 would result in a zero-beta combination. This follows since the liability is an outflow and thus is a negative 1.5 beta. The negative 1.5 beta and the positive 1.5 beta is a zero-beta combination insensitive to market movements.

To clarify further, consider a single liability of $100 at year 5. The goal of the investment program is to meet that liability. If a bond is purchased with a maturity of five years, the investor is certain about the value of the bond at the horizon but is uncertain about the rate at which coupon payments will be invested. If interest rates rise, the obligation will be more than met because the coupon payments will be invested at rates that were higher than anticipated. However, if interest rates fall, the obligation will not be met because the coupon payments will be invested at a rate below what was anticipated. If the investor purchases a bond with a maturity of longer than five years, the investor will also be uncertain about the value of the bond at year 5. Consider a rise in interest rates. With a rise in interest rates, the aggregate value of the coupons at the horizon will be higher than anticipated because of the coupon payments being invested at more favorable rates. However, because interest rates rose, the value of the bond at the horizon will be less. These influences work in opposite directions. If the bond is selected properly, these effects will exactly balance one another. Similarly, consider a decline in interest rates. With a decline, the coupon payments will be invested at rates less than anticipated. The aggregate value of the interest payments at the horizon will be less. However, if interest rates decline, the value of the bond will rise. Once again, it might be possible to choose a maturity so that these influences exactly offset one another.

Table 22.4 The Value of a Bond with Changing Interest Rates

		Value as of Period 4		
Time	Cash Flow	11%	10%	12%
1	13.52	$13.52(1.11)^3$	$13.52(1.1)^3$	$13.52(1.12)^3$
2	13.52	$13.52(1.11)^2$	$13.52(1.1)^2$	$13.52(1.12)^2$
3	13.52	$13.52(1.11)^1$	$13.52(1.1)^1$	$13.52(1.12)^1$
4	13.52	13.52	13.52	13.52
5	113.52	$113.52(1.11)^{-1}$	$113.52(1.1)^{-1}$	$113.52(1.12)^{-1}$
		165.946	165.946	165.974

The principles discussed in this part of the chapter are exactly why immunization works. At a point in time equal to the duration of the assets, the change in reinvestment income will exactly match the change in the value of the bonds. Table 22.4 illustrates these ideas. Assume that interest rates are currently at 11% for all maturities. Further assume that the bond pays annual interest of 13.52% and has a maturity of five years. These are the flows shown in the second column of Table 22.4. The duration of this bond is four years. The value of this bond as of period 4 if interest rates remain at 11% is 165.946.

If interest rates decline to 10%, the value as of period 4 is 165.946. The value is unchanged because the decrease in the value of the interest payments of 0.930 is exactly offset by an increase in the value as of period 4 of a payment of 113.52 in period 5. This increase is 0.930. If interest rates rise to 12%, the value of the coupon payments as of period 4 increases, while the value as of period 4 of receiving 113.52 at period 5 decreases. Although these do not completely offset one another, they come close to doing so. This example illustrates the idea of immunization. If we had a liability at period 4, we could purchase a sufficient quantity of the bond to just meet the liability. For example, a $995 liability could be met with six bonds. Whether interest rates decrease or increase, the same liability could be met.

Why does the bond in Table 22.4 have these properties? The coupon for the bond in Table 22.4 was selected so that the bond has a duration of four years. Pure discount bonds have a duration equal to their maturity. Thus a pure discount bond with a maturity of four years also has a duration of four years. Earlier we argued that duration is a measure of sensitivity to interest rate changes. Two bonds with the same sensitivity have their value change by the same amount. If one bond could be swapped for a second before an interest change, it could also be swapped after the change. Because the pure discount bond has a constant value as of period 4, the bond that could be swapped for it would also have a constant value as of period 4.

In the last section we discussed how the addition of convexity improved the approximation of the estimated price change to the true price change. Many managers engaging in immunization match on convexity as well as duration. Their concern is that the convexity of the liabilities and assets might be quite different and the approximation utilizing duration alone might lead to large errors. The addition of convexity involves a trade-off. The addition of convexity should provide better protection against term structure shifts. However, fewer portfolios will be both duration and convexity matched. Thus the match on both measures will likely result in higher cost portfolio.

Immunization strategies are widely used to mitigate the effect of interest rate changes. Extensive research has been done on designing immunized portfolios. We now discuss some implications of this research. The duration on a portfolio of bonds is a weighted average of the duration of the individual assets that make up the

portfolio.[7] Let X_i be the proportion of bond i in the portfolio, D_i be the duration of asset i, and D_P be the duration on the portfolio with N bonds:

$$D_P = \sum_{i=1}^{N} X_i D_i$$

There are obviously an enormous number of ways to construct a portfolio of a particular duration. For example, assume that a bond portfolio with a duration of 10 years is required. Further assume that four bonds are being considered with a duration of 6, 8, 10, and 12 years. Simply holding the bond with a duration of 10 years would meet the constraint. Alternatively, one-sixth of the money could be invested in the bond with 6 years' duration, one-fourth in the bond with 8 years' duration, and the remaining seven-twelfths in the 12-year bond. This results in a duration of 10 years because

$$\left(\frac{1}{6}\right)6 + \left(\frac{1}{4}\right)8 + \left(\frac{7}{12}\right)12 = 10$$

Two different strategies have been explored: a barbell strategy and a focused strategy. The focused strategy finds a portfolio of bonds with each bond having a duration close to the duration of the liability. For example, if the liability is 10 years, then the bonds might have a duration between 9 and 11 years. The bond portfolio is focused around the duration of the liability. The barbell strategy uses bonds with very different durations, for example, 5 and 15 years. The 10-year duration would be met by one-half in the 5-year duration bonds and one-half in the 15-year duration bonds. The advantage of a barbell strategy is that there is no necessity to construct individual bond portfolios to meet each liability. Instead, liabilities of different duration can be met by selecting different mixtures of the 5- and 15-year duration portfolios.

These two strategies have been explored to determine which one better meets the goal of having the asset and liability mix equally sensitive to changes in interest rates. The empirical evidence gives some support to the focused strategy. The reason seems to be as follows. All duration measures are approximations of the effect of the true shift in interest rate patterns. When individual assets and liabilities have similar durations, these errors are similar. When the individual assets in a portfolio have different durations from the liabilities even though the portfolio has the same duration, the error patterns can be very different. This latter pattern is what occurs with a barbell strategy. Thus inaccuracies in the duration estimate explain in part the evidence tending to support focusing.

Before closing this discussion, one more facet of immunization should be discussed. Immunization is often presented as a passive strategy, and therefore one by which a set of bonds is purchased and held to maturity. This impression is incorrect. Duration is calculated for a particular yield curve. As the yield curve shifts, duration changes and the assets and liabilities may no longer have the same duration. If the differences become large enough, restructuring is required. Furthermore, even if the yield curve stays constant, the duration of the assets and the liabilities will move apart, unless both assets and liabilities have the same cash flow pattern. This also requires restructuring. Thus immunization is an active strategy.

What are the risks of an immunized strategy? The principal one is the selection of the wrong duration measure. Each duration measure is derived assuming a different pattern of

[7]This is a property of most duration measures. For the duration measure discussed here, it holds only if the yield curve is flat.

shifts in the yield curve. A portfolio is usually immunized using one measure but not another. For example, using a measure that accurately measures price change for parallel shifts in a flat yield curve will not accurately measure price change if the yield curve steepens (long rates increase more than short).

Lest the reader become overly concerned, it is worth repeating that even the simplest measure discussed in this chapter works very well. The second risk of immunization concerns major yield shifts when the portfolio is not immunized. As discussed previously, either the passage of time or small changes in the yield curve will result in the portfolio not being immunized. Cash flows from the portfolio are used to purchase bonds to rebalance the portfolio so the duration of assets is closer to the duration of liabilities. Bond sales and purchases (rebalancing) could also be used to immunize the portfolio exactly. However, bond swaps are costly, so that a manager will let the duration of the assets drift away from the duration of the liabilities and not be immunized at all points in time.[8] The risk is that just before the manager engages in a bond swap to adjust the duration to immunize the portfolio, the interest rates may change dramatically.

A cash flow matched portfolio is, of course, immunized. Because its immunization comes from matched cash flows rather than the accuracy of a measure, it is generally less risky. Thus the immunized portfolio has to be less costly than the cash flow matched portfolio for an organization to immunize. It is often optional to cash flow match part of the portfolio and immunize the remainder.

In this section, we have presented techniques for protecting against interest rate shifts. In the next section we discuss techniques for constructing portfolios when performance over a one-year period is being evaluated.

BOND PORTFOLIO MANAGEMENT OF YEARLY RETURNS

In the prior sections we have discussed designing portfolios of bonds that are reasonably insensitive to changes in the yield curve. The return on these portfolios can fluctuate dramatically from period to period because the concern is meeting some future liability rather than period-by-period returns. Many managers are interested not in meeting some future liability but in the year-by-year return on the portfolio. Managers of bond funds and many managers of pension funds are concerned with year-by-year variability.

This section is divided into three parts. In the first part we discuss indexation. Indexation is the passive strategy used by managers interested in period-by-period returns. The second section discusses active bond management techniques. In the third part, we discuss active portfolio management using modern portfolio theory.

Indexation

Another passive strategy finding favor with bond managers is index replication. The major motivation behind index replication in the bond area is performance. Very few actively managed funds have outperformed the major bond indexes. Given this experience, many pension managers, in particular, have indexed a part of their assets. Indexation in the bond area is done differently than in the common stock area. There are thousands of corporate bonds, many of which are completely inactive. Thus holding bonds in the same proportion as the index is infeasible. Rather, indexation is commonly done via cell matching. The major important characteristics of a bond are delineated. These generally include category

[8]Futures can be used to adjust duration (see Chapter 24) as well as interest rate swaps. These are generally less expensive alternatives than bond swaps.

(government, corporate, utility, etc.), duration, coupon, and bond rating. Then the proportion of the index with any set of characteristics is determined over all possible characteristics. For example, what percentage of the index is represented by corporate bonds rated BAA, with a duration between four and five years and a coupon between 8% and 9%? These percentages are calculated for all possible combinations of bond characteristics. A portfolio of bonds is then constructed that has roughly the same proportion in each cell as the indexes. This type of index replication is highly successful in matching the performance of the index.

Active Bond Management

There are essentially four categories of active investment strategies in the bond area. These are aggregate interest rate forecasting, sector selection or rotation, and individual bond selection.

Aggregate Interest Rate Forecasting The major cause of variation in year-to-year return for a manager is unexpected shifts in the yield curve. Examining Table 22.1 shows that for most years the unanticipated return was considerably larger in absolute magnitude than the anticipated return. For example, in 1993, long-term bonds returned 16.38% while intermediate-term bonds returned 7.91%, even though the expected return on intermediate- and long-term bonds would have been very similar. Likewise, in the 1980s, the long bond return varied from about −3% to +42%, with much less variation in expected return. We know from our discussion of duration that if interest rates rise unexpectedly, short-duration bonds will be hurt less than long-duration bonds, and if interest rates fall unexpectedly, short-duration bonds will gain less than long-duration bonds.

Thus one investment strategy that managers follow is to shorten the duration when they expect rates to rise more than is anticipated by the market (and reflected in the yield curve) and lengthen the duration when rates are expected to fall more than anticipated by the market. Bond managers pay a price for this timing. Most bonds are not as liquid as common equities. Those bonds that have a large market and can be readily traded in a short period of time are primarily government bonds of certain special maturities. Restricting purchases to these bonds can result in a lower expected return compared to purchasing corporate bonds with higher expected returns or bonds that are mispriced. In addition, folklore and possibly empirical evidence suggest that the bonds used in timing have a lower return than comparable risk bonds because of their marketability. Finally, concentrating on a few government issues to facilitate timing results in a relatively undiversified portfolio.[9]

No forecaster is accurate all of the time. For a forecast of future interest rates to be useful, it has to be accurate and different from the consensus, because the consensus is already reflected in existing rates. A forecaster should be correct in estimating whether interest rates will rise or fall 50% of the time by chance. A forecaster who is accurate 60% of the time in calling direction would be doing extremely well in forecasting the market. Market timing involves one estimate each period: future interest rates. Because even a good forecaster will often be wrong, it will take a number of periods before there is a high probability that a manager with timing ability has superior returns.

[9]An alternative technique for changing duration is the use of futures (see the discussion in Chapter 24). Using futures to change duration has two advantages: it is cheaper and it allows a separation of the selection decision and the duration decision. Thus the manager selects the cheapest bonds and manages duration by using futures.

Some managers immunize and also engage in some market timing. For these managers the normal strategy is to have the duration of the assets and liabilities the same. If they anticipate rates to rise more than the market expects, then the duration of the assets is set less than the liabilities, and if they expect rates to fall more than the market expects, the duration of the assets is set greater than the duration of the liabilities. Managers immunizing the portfolio are likely to be very cautious in utilizing market timing.

Sector Selection Managers who engage in sector selection are doing so because they believe that in the long run, some sector will give superior performance. The most common type of sector selection is to lower the average credit rating on the portfolio. For example, a manager could believe that junk bonds offer a larger risk premium than is justified by any difference in risk and permanently invest in junk bonds in the belief that in the long run, this premium will be earned and junk bonds will outperform other categories.

Sector Rotation Sector rotation can be practiced using any of the characteristics of bonds discussed earlier. Sector rotation is related to sector selection. Sector rotation involves overweighting a sector in the belief that the relative performance of this sector will be better in the *next period.* For example, the yield to maturity on AAA corporates selling at par is higher than the yield to maturity on governments of the same maturity selling at par. This difference is partially a default premium and partially a risk premium. The spread would widen if investors believed that default risk increased. If a manager believed that the market was overreacting to a perceived risk increase, then the manager would switch to AAA debt. If the manager's assessment was correct, the manager would earn a larger than normal risk premium in the period and could earn an additional return if the default premium subsequently narrowed because of many investors realizing they have overreacted. For example, assume the normal default premium between government and corporates was $\frac{1}{8}$% and the default premium widened to $\frac{1}{4}$%. If the spread goes back to an $\frac{1}{8}$%, then the yield on AAA corporates is falling relative to Treasuries, and AAA corporate prices will rise relative to governments.

As mentioned earlier, sector rotation can be practiced with respect to any of the factors affecting bond prices. As a second example, assume the investor feels the market is underestimating the volatility of interest rates. The more volatile the interest rates, the greater the change in interest rates that can occur. The greater the change in interest rates, the more likely very low interest rates will occur, and it will pay a firm to call a bond. Thus an investor believing that the market has underestimated volatility will believe that callable bonds are relatively unattractive and will rotate away from callable bonds.

Mispriced Bonds There are generally two procedures for bond security selection. One is to accept bond classifications as accurate (e.g., AAA or AA) and to try to find the most attractive bonds in a given class. The second procedure is to look for misclassified bonds. For example, a firm might treat all AA noncallable bonds with 8 to 10 years' maturity as equivalent with respect to risk. The firm could then examine all bonds that met this criteria and select the most attractive. Brokerage firms generally utilize yield to maturity as a metric of desirability. Thus, they would suggest bonds with the highest yields as the most desirable.[10] Bond services such as Barra or Gifford Fong utilize the difference between actual price and theoretical price as a metric of desirability. Theoretical price is determined by discounting future cash flows at estimated spot rates and adjusting the price for any option value.

[10]For some bonds such as Ginnie Maes, firms use spread over comparable Treasuries as a measure of desirability. We have discussed the difficulties with yield measures in the prior chapter.

The other way firms practice bond selection is to look for misclassified bonds. This is especially prevalent with low-rated bonds. Implicit in the bonds' rating is a default probability and expected loss in event of default. The firm practicing this method of selection examines the issuing firms' characteristics and tries to find bonds that have default probabilities or expected loss that is different than what is implied by the bond ratings. Those with more attractive characteristics are selected.

In the next section we discuss techniques for selecting bonds similar to those used for stocks.

Active Bond Selection Using Modern Portfolio Theory

Modern portfolio theory can be applied to bond management as well as stock management. In this section we discuss how this can be done.

Estimating Expected Return We start this section with a consideration of the simplest class of bonds: noncallable bonds issued by the federal government. Later we discuss expected return on nongovernment bonds and the impact of callability and tax considerations.

While any of the theories of the term structure of interest rates can be used to estimate the expected returns on a bond, let us start off illustrating the methodology with the simplest term structure theory: the expectations theory. Under the expectations theory, all bonds must give the same rate of return over any specific time horizon. Thus next period's expected return for any bond is simply the one-period spot rate.

This is modified if we recognize that bonds may be mispriced. Then the expected return will be a function of mispricing if it exists as well as the one-period spot rate. To see the impact of this on expected returns, it is necessary to make an assumption about the period of time that elapses before the market corrects mispricing. We will follow common practice and assume that prices adjust to equilibrium within one period. This is the assumption implicit in most commercial services and seems consistent with empirical evidence. To calculate the expected rate of return on a bond, we need to calculate its expected equilibrium value one period in the future. Then, from the interest payment expected during the period and the expected capital gain (change in price), we can calculate the expected rate of return.

To get a price for a bond one period in the future, we need expectations about what spot rates or forward rates will be at that time. In the previous chapter we showed how to derive forward rates from the spot rates. If the expectation theory holds, forward rates are not expected to change over time. A hypothetical set of rates is shown in Table 22.5. Assuming these rates, let us examine the expected return on a bond that will mature in five years and pays interest of $8 per period. The bond has $100 principal payment, and its current price is $82.

Table 22.5 Hypothetical Set of Rates

Period	Current One-Period Forward Rate (%)	Expected Forward Rate in One Period (%)
1	10	
2	11	11
3	12	12
4	13	13
5	14	14

If the bond were priced in equilibrium at the initial period, its price would be

$$P_0 = \frac{8}{1.10} + \frac{8}{(1.10)(1.11)} + \frac{8}{(1.10)(1.11)(1.12)} + \frac{8}{(1.10)(1.11)(1.12)(1.13)} + \frac{108}{(1.10)(1.11)(1.12)(1.13)(1.14)}$$

$$= \$86.16$$

The expected price one period in the future is

$$P_1 = \frac{8}{1.11} + \frac{8}{(1.11)(1.12)} + \frac{8}{(1.11)(1.12)(1.13)} + \frac{108}{(1.11)(1.12)(1.13)(1.14)}$$

$$P_i = \$86.77$$

Note that if the bond had been priced in equilibrium at time 0, the one-period cash flow would have been \$8 in interest and 61¢ in capital gains, for a total return of 8.61/86.16 or 10%. The 10% is, of course, the spot rate in the first period. If instead the bond could have been bought for \$82, the return would be \$8 in interest and \$4.77 in capital gains or a return of 15.57%. This rate of return can be broken into its three components: 9.76% from interest income, 0.74% from the change in the equilibrium value of the bond, and 5.07% from the effect of mispricing.

If an alternative term structure theory is a better description of reality, there is a further element to expected return. However, the same techniques are applicable even if any of the other alternative term structure theories is a better description of reality. Consider the liquidity premium theory as an example. With the liquidity premium theory, the expected return is the one-period spot rate plus any adjustment so that the bond is priced in equilibrium plus the change in the liquidity premium. The same procedure can be used to value bonds as was discussed with the expectation theory, but the effect of the change in the liquidity premium has to be taken into account. Consider the example shown in Table 22.6.

Table 22.6 is divided into two parts: calculations associated with the current period and calculations associated with one period in the future. The table shows forward rates in the current period. These one-period rates can be determined from spot rates using the techniques discussed in the prior chapter. In the third column is a set of hypothesized liquidity premiums. These are subtracted from the forward rates to arrive at the forward rates without the liquidity premium shown in the fourth column. These rates are assumed to remain unchanged. Thus the fifth column is the same as the fourth column. The column that is

Table 22.6 Assumed Forward Rates (in Percentages)

	Current Period			Next Period		
Period	Forward Rates	Liquidity Premium	Forward Rate (Liquidity Premium Removed)	Forward Rate (Liquidity Premium Removed)	Liquidity Premium	Forward Rates
1	10		10			
2	11	0.1	10.9	10.9		10.9
3	12	0.2	11.8	11.8	0.1	11.9
4	13	0.3	12.7	12.7	0.2	12.9
5	14	0.4	13.6	13.6	0.3	13.9

changed is the liquidity premium column. The liquidity premiums are the same; however, each premium is moved one period in the future. Thus the 0.1 liquidity premium that was the premium for two-year money as of the initial period appears in the third period rather than the second period because at time 1, two periods in the future are period 3.

Assume the same bond discussed previously: a bond with an 8% coupon and a $100 principal payment. Further assume that it sells for its equilibrium price. The forward rates shown in Table 22.6 as of the current period are identical to the rates in Table 22.5, and thus the equilibrium price is unchanged or

$$P_0 = \$86.16$$

The equilibrium price in one period using the rates shown in Table 22.6 is

$$\begin{aligned} P_1 &= \frac{8}{1.109} + \frac{8}{(1.109)(1.119)} + \frac{8}{(1.109)(1.119)(1.129)} \\ &\quad + \frac{108}{(1.109)(1.119)(1.129)(1.139)} \\ &= \$87.05 \end{aligned}$$

Without an assumption of a liquidity premium, the equilibrium price in period 1 was $86.77. The difference between $87.05 and $86.77 is the effect of the additional capital gain due to bearing maturity risk. Total expected cash flow is interest income of $8, an expected capital appreciation without the liquidity premium of $86.77 − $86.16 or 61¢, and an effect of the liquidity premium applying to different cash flows of $87.05 − $86.77 or 28¢. Total expected return is (8 + 0.61 + 0.28) divided by $86.16 or 10.32%. The extra 0.32% is the liquidity premium effect. Any mispricing can be dealt with as discussed earlier for the expectations theory.

Up to now we have ignored the effect of default risk, callability, or tax effects in this discussion. Although there are many ways to deal with these influences, we will briefly discuss what has become the most widely used technique. To keep the discussion simple, we will assume the expectations theory holds, though the modifications for the liquidity premium theory are straightforward and follow from the discussion of how to deal with the liquidity premium presented before.

Let us look at callability. The future prices for noncallable government bonds are arrived at by the prior methods. Prices for callable bonds are arrived at by using the rates for noncallable bonds. The average difference between actual price for all callable bonds and the price arrived at for these bonds when they are priced as if they were noncallable bonds is then calculated. The theoretical price of any callable bond is arrived at by pricing as if it were a noncallable government bond and adding the average difference. Mispricing is the difference between the actual price and this theoretical price. This is obviously a crude procedure. A much more exact procedure would use the option pricing models of Chapter 23 to arrive at an estimate of the differential price due to callability. This differential price would be used to estimate possible mispricing. Taxes and default risk are evaluated in an analogous manner.[11]

Index Models In Chapters 7 and 8 we discussed methods of estimating the variance–covariance structure of common stock returns. The general principles discussed are

[11]As an alternative to this procedure, some managers estimate spot rates and the effect of callability, default risk, and taxes simultaneously using a multiple regression and the techniques discussed in Appendix A at the end of the chapter. Once again, an assumption is made that spot and forward rates remain unchanged, and a new price is estimated one period in the future. This new price is used to calculate an expected return.

equally as applicable to bonds as they are to stocks. However, there are special characteristics of bonds that suggest that some modification and respecification would be useful.

Single-Index Models In this section, we discuss the application of the single-index model to bond portfolio management. Consider first applying it to noncallable government bonds with no special tax effects. The return on government bonds can be divided into two parts: the anticipated return and the unanticipated return due to both changes in the yield structure and/or changes in the pricing of the bond in question relative to the yield structure. As discussed previously, if the expectations theory is correct and bonds are fairly priced, then all bonds should have the same expected return over the first period. If one of the other theories is correct or there is mispricing, then the bonds may have different returns, and these returns will depend on the maturity of the bond. We will derive the single-index model under the assumption that the expectations theory holds.

The unanticipated return has two sources: a change in the yield curve or a change in the bond price relative to the yield curve. In the first section we showed that the return on a bond due to a shift in the yield curve was minus duration times a measure of interest rate change. We also emphasized that the duration measure is based on a simplified assumption about unanticipated shifts in the yield curve. Assume that the influence of shifts other than that assumed in deriving the duration measure is random. Further assume that shifts in the bond return relative to the yield curve are random. With these assumptions, the effect of these two influences on return are random and can be represented by e_i, where the expected value of e_i is zero and the variance of e_i is represented by σ_{ei}^2.

Let us put these ideas together as follows:

$$\begin{array}{c}\text{Total}\\ \text{return}\end{array} = \begin{array}{c}\text{Expected}\\ \text{return}\end{array} + \begin{array}{c}\text{Return due to an}\\ \text{unanticipated shift}\\ \text{in the yield curve}\end{array} + \begin{array}{c}\text{Random}\\ \text{influence}\\ \text{on return}\end{array}$$

$$R_i = \bar{R}_i - D_i\,\Delta + e_i \tag{22.5a}$$

where

- R_i is the return on bond i
- $\bar{R}_i$ is the expected return of bond i
- D_i is the duration of bond i
- Δ is the change in interest rate divided by 1 plus the interest rate
- e_i is the random influence with a mean of zero and a variance of σ_{ei}^2

In Chapter 7 we expressed the single-index model in terms of an equity index. We can express the return on a bond in terms of a bond index. Let X_i^m be the proportion of bond i in the bond index. Then the return on the index called R_m is

$$\begin{aligned} R_m &= \sum_i X_i^m R_i = \sum_i X_i \bar{R}_i - \sum_i X_i D_i(\Delta) + \sum_i X_i e_i \\ &= \bar{R}_m - \sum_i X_i D_i \Delta + \sum_i X_i e_i \end{aligned}$$

For a bond index with a large number of bonds, $\Sigma_i X_i e_i$ should be approximately zero. This follows from assuming that the e_i are independent from one another. Define D_m as $\Sigma_i X_i D_i$ or the duration of the bond index. With these substitutions, we have

$$R_m = \bar{R}_m - D_m(\Delta) \tag{22.5b}$$

Solving (22.5b) for Δ and substitution into (22.5a) yields[12]

$$R_i = \overline{R}_i + \frac{D_i}{D_m}\left(R_m - \overline{R}_m\right) + e_i \tag{22.6}$$

To complete the analogy with the model discussed in Chapter 7, define β_i as D_i/D_m. With the assumption of e_i being independent of the bond index, β_i has the same meaning as in Chapter 7, that is, β_i is the covariance of R_i with R_m divided by the variance of R_m. However, there is no reason to estimate β_i using historical or modified historical data. Instead, it can be measured directly as the ratio of durations.

Equation (22.6) is analogous to the single-index model presented for common stocks. If we make the assumption of the single-index model that $E(e_i e_j) = 0$ for $i \neq j$, then we find

$$\text{cov}\left(R_i R_j\right) = \frac{D_i D_j}{D_m^2}\sigma_m^2$$

$$\text{var}\left(R_i\right) = \frac{D_i^2}{D_m^2}\sigma_m^2$$

This is not surprising because, as we have already stated, $\beta_i = D_i/D_m$. Single-index models have been used widely in stock selection. There is much less experience concerning their usefulness in the bond management area. Single-index models for bonds did not appear commercially until the 1980s. Similarly, there has been very little academic research into the applicability of single-index models to bond management. This is in contrast to the extensive research done in the common equity area.

Before leaving this discussion, we want to mention some other influences affecting returns on bonds. These include liquidity premiums, tax effects, callability, and default risk. If the impact of all of these influences were constant over time, then the single-index model would be appropriate. However, if the premium for these influences changed over time, then bonds would have an added source of variance and covariance. These added influences might be an added source of covariance, just as industry membership might be for common stocks. For example, two AAA-rated bonds might move more alike than two bonds picked at random. This leads us logically to the next section of this chapter, on multi-index models.

Multi-index Models There are a number of reasons why a multi-index model might be more relevant than a single-index model (several were discussed in the last section). The major reasons are as follows:

1. to more accurately measure the effect of interest rate changes
2. to reflect the variability introduced by the change in the yield spread between bonds of a particular risk class and governments
3. to reflect the variability introduced by the change in yield spread between bonds from various sectors: government, financial, and corporate
4. to reflect the variability introduced by the change in the value of a call
5. to reflect the variability introduced by changes in the importance of taxes

---

[12]Using arbitrage pricing theory, this return-generating process results in the following equilibrium model:

$$\overline{R}_i = \overline{R}_Z + \frac{D_i}{D_m}\left(\overline{R}_m - \overline{R}_Z\right)$$

Any of these influences could be important enough so that a multi-index model would reflect the covariance structure better than a single-index model.

A number of studies have shown that two factors are necessary to capture changes in the term structure.[13] An example of the two factors researchers have used is changes in the long rate and changes in the spread between the long and short rate. Consider, for example, the following two-factor model:

$$R_{it} = \overline{R}_i + \beta_{i1}F_{1t} + \beta_{i2}F_{2t} + e_{it}$$

where

R_{it} is the return on bond i in period t
$\overline{R}_i$ is expected return on bond i
β_{ij} is the sensitivity of bond i to factor j
F_{jt} is the value of factor j in period t
e_{it} is the random error term

Using two factors seems to substantially improve the explanatory power of these types of return-generating processes. To be more concrete concerning the factors, consider an example. As a proxy for the long rate, some investigators have used the rate on a 10-year government bond. Factor 1 in period t would be the change in the interest rate on a 10-year government bond from period t to period $t + 1$. The change in the interest rate measures the shift up or down in the term structure. One would expect β_{ij} to be negative so that if interest rates increased, the price on the bond would decline and the unexpected part of return due to an upward shift in the yield curve would be negative. Some investigators use a change in the short rate for factor 2; others use changes in the spread between long and short bonds as the second factor. For example, the spread could be the difference between 10-year and 1-year rates. The change in the spread between these rates from period t to $t + 1$ would be the value of the second factor. An increase in the spread between long and short rates while holding long rates constant implies a decrease in short rates. This should result in positive return for short bonds; thus b_{i2} should be positive.

Estimating the sensitivities in a return-generating process for bonds is more difficult than it is for common equities. In common equities, a time series regression of return on factors is the usual starting point for most estimations. With bonds, the maturity shortens as time passes. It is generally believed that sensitivity is related to maturity. For example, in the one-factor model, when sensitivity was related to duration, as maturity shortened, so did duration, and hence sensitivity changed. For time series regression to be an appropriate method of estimating sensitivity, the sensitivity must remain constant over time. Thus time series estimation of sensitivity for individual bonds is probably inappropriate.

What has been done is to estimate the sensitivities for a pure discount bond of constant maturity. Because coupon-paying bonds can be viewed as portfolios of pure discount bonds and because the sensitivity on a portfolio is a weighted average of the sensitivity of the bonds comprising it, this procedure can be used to estimate a bond's sensitivity. For example, each month the return is calculated on the factors and on a 10-year pure discount bond. Of course, the bond that is a 10-year pure discount bond changes each month. The sensitivities are then estimated by regressing the return on the 10-year pure discount bond on the two factors. Any coupon-paying bond can be viewed as a portfolio of pure discount

[13]See, for example, Brennan and Schwartz (1983), Nelson and Schaefer (1983), Elton, Gruber, and Naber (1988), and Elton, Gruber, and Michaely (1990).

bonds. The sensitivity on a portfolio is a weighted average of the sensitivities of the components, where the weights are the proportion each component represents of the whole. For example, define

1. b_{t1}, b_{t2} as the sensitivities of a t-period pure discount bond to factor 1 and 2, respectively
2. $PV(Cf_{ti})$ as the present value of the cash flow for bond i in period t
3. P_i as the price of bond i

The sensitivities for bond i are a weighted average of the sensitivities on the pure discount bonds, or

$$\beta_{i1} = \sum_t \frac{PV(Cf_{ti})}{P_i} b_{t1}$$

$$\beta_{i2} = \sum_t \frac{PV(Cf_{ti})}{P_i} b_{t2}$$

There are other ways to estimate the sensitivities. For the one-factor model, we could derive, using duration, a theoretical value for the sensitivities. There are two parameter duration models that allow a similar derivation of the sensitivities for two-factor models. Finally, other researchers have used duration for the first factor and convexity for the second. Both of these factors can be directly calculated.

Commercially available bond models generally estimate yield on bonds (and the corresponding price) rather than the period-by-period returns. Popular examples of these models are those sold by Barra and Fong. These models generally are multifactor models. They usually have two-term structure terms and additional terms to capture the spread between corporates and governments and option features of the bonds. These models can be used in two ways. One way is to try to select individual bonds that are mispriced in the sense that the model and theoretical price diverge. A second use of the models is to control the sensitivity to the factor. If one believes the spread between longs and shorts is going to change, then one could adjust the sensitivity to spread accordingly. Finally, these models can be utilized to estimate risk for portfolio purposes. The discussion in Chapters 7 and 8 of how to use index models for portfolio risk estimation is equally applicable to models for returns on bonds.

SWAPS

In recent years, swaps have become an increasingly important part of bond management. Bond managers can swap bonds, or they can swap interest rate streams. We discuss each in turn.

Bond Swaps

Bond swaps are divided into several categories based on the purpose of the swap. We discuss the major categories.

Substitution Swap The substitution swap is a swap of two bonds that are identical in characteristics but have different prices. Assume two 10-year government bonds; both have coupons of 8%, and one has a lower price than the second. A substitution swap is selling the higher-priced bond and buying the lower-priced bond. In Appendix C, we generalize a substitution swap to the case where a portfolio is being swapped for a second portfolio

with the same cash flows and bond characteristics. For the existence of a substitution swap, there must be a violation of the law of one price. Profitable substitution swaps are likely to be rare when one bond is being swapped for a second bond. They are more likely to exist when they involve complex combinations of large numbers of bonds.

Yield Pickup Swaps The yield pickup swap is swapping a bond with a lower yield to maturity for a bond of like risk and maturity but a higher yield to maturity. As we discussed in the last chapter, yield to maturity on a portfolio is not a weighted average of the yield to maturity of the bonds that comprise it. Thus swapping one bond for a higher-yield bond can actually reduce the yield on the portfolio. Furthermore, the bond with the higher yield to maturity could be overpriced when price is determined by discounting the cash flows at the spot rates, while the lower yield to maturity bond is fairly priced. Thus, although yield pickup swaps are frequently discussed, the logic underlying them is tenuous.

Tax Swaps Individuals in many countries including the United States are subject to tax on realized capital gains and losses. A tax swap involves generating a capital loss to offset either capital gains or, to a limited extent, ordinary income. Assume an investor has a bond that is selling for less than it was purchased but wishes to hold a security with the same characteristics as that bond. The investor can sell the bond whose value has declined, generating a capital loss, and purchase a bond with identical characteristics.[14] This action is a tax swap.

In the United States the Internal Revenue Service will not allow an individual to claim a capital loss if the purchase and sale involve the same security (wash sale). With bonds, however, it is usually easy to find a second bond that is almost identical to the first in coupon, maturity, and risk. Tax swaps are especially advantageous with municipal bonds. Assume an investor holds municipal bonds and interest rates rise. The investor's bonds fall in value. Because interest is not taxable on municipals (at least at the federal level), a sale of the municipals that declined in price and a purchase of a similar new municipal at par results in a capital loss with no corresponding tax obligation on the purchased bond.

Interest Rate Swaps

One of the major investment tools used in fixed-income management is the interest rate swap. Interest rate swaps involve exchanging interest streams without exchanging the securities. The most basic type of swap is the fixed-for-variable swap. In this type of swap, one party agrees to pay the other party a fixed coupon in return for a variable coupon. For example, party A might agree to pay a 6% coupon semiannually to party B over the next five years in return for a variable coupon equal to the six-month Treasury bill rate that exists at the beginning of each six-month period. The two parties have to agree not only on the rate but also on the principal amount to which the interest rate is applied, called the *notational principal*. If the notational principal was $10 million, then the flows would be as depicted in Table 22.7.

Interest rate swaps are arranged by all the major brokerage firms. The parties engaged in the swap may or may not know who is on the other side. Swaps are an alternative to a

[14]This swap may generate a capital gain in the future. For example, assume the investor bought the bond at $100 and it declined to $80. Selling the bond generates a $20 capital loss. Now assume the investor buys an identical bond at $80. When the bond matures at $100, the investor would need to pay a $20 capital gain. If this is in a subsequent tax year, the present value will be less than $20, and the investor will gain.

Table 22.7 Cash Flows of a Fixed-for-Floating Swap Assuming a $10 Million Notational Principal

Time Period (in half-years)	Six-Month T-bill Rate (Annually)	Paid by B to A	Paid by A to B
1	6%	$300,000	$300,000
2	4%	$200,000	$300,000
3	7%	$350,000	$300,000
4	6%	$300,000	$300,000
5	5%	$250,000	$300,000
6	7%	$350,000	$300,000

direct sale. An investor interested in exchanging a long-term security for a sequence of six-month T-bills could potentially sell the long-term security and buy a series of six-month T-bills. Although it does not involve a physical sale, the swap serves the same purpose. Why the swap?

First, swaps are relatively inexpensive.[15] Thus it may be cheaper to swap interest rate streams rather than sell a long-term bond and purchase a short-term bond. Second, one or more of the parties may not wish to sell the asset. For example, savings and loans hold mostly long-term mortgages on properties in their local community as their assets. One of their major liabilities is short-term savings accounts. To protect against term structure shifts, they would like to have the duration of the assets and liabilities matched. The savings and loan may feel that to maintain local goodwill, they need to hold the long-term mortgages. A fixed-for-floating swap can be used to duration match without physically selling off the assets. A third reason for a swap is comparative advantage. It has been argued that the risk premium that low-quality firms have to pay in issuing fixed debt is higher than they have to pay for variable rate debt. Furthermore, because the interest rate swap does not involve the principal, only the interest stream, bankruptcy of one of the parties can only cost the other the opportunity cost of not having a favorable interest rate exchange.[16] Thus it is argued that high-rated and low-rated corporations could gain by a swap. The swap involves a high-rated corporation wanting to borrow at a variable interest rate instead borrowing long at a fixed rate, and then swapping fixed for floating with a low-rated corporation that wants to borrow fixed.

APPENDIX A

DURATION MEASURES

There are at least a dozen different measures of duration. Duration measures the sensitivity of bond prices to a change in the yield curve. In the text we assumed that the yield curve was flat and there was a parallel shift in the yield curve. A large number of alternative assumptions are possible. The yield curve could be upward or downward sloping, and the shift could be very different than parallel. Each of these alternative definitions results in a different measure of duration. In the text we derive one measure

[15]Estimates that the bid–ask spread is about 5 basis points.

[16]The fixed-for-variable swap is the most common swap. Other types of swaps involve interest rate swaps in different currency (used to manage currency risk) and floating-rate swaps where the floating is tied to different instruments.

of duration. This is the measure most often used. The second most common is derived as follows.

1. Macaulay's Second Measure

Assume that the yield curve is not flat but that spot rates vary. Let S_{0t} be the spot rate for a t-year bond. Consider a pure discount bond that pays \$1,000 at year t. Its price is

$$P_0^t = \frac{1{,}000}{(1+S_{0t})^t} = 1{,}000(1+S_{0t})^{-t}$$

Its sensitivity to a change in $1 + S_{0t}$ is

$$\begin{aligned} dP_0^t &= 1{,}000(-t)(1+S_{0t})^{-t-1}\, d(1+S_{0t}) \\ &= \frac{1{,}000}{(1+S_{0t})^t}(-t)\frac{d(1+S_{0t})}{1+S_{0t}} \end{aligned}$$

Recalling that $P_0^t = 1{,}000/(1 + S_{0t})^t$ and dividing through by P_0 yields

$$\frac{dP_0^t}{P_0^t} = -t\frac{d(1+S_{0t})}{(1+S_{0t})}$$

The key assumption of the second measure of duration is that the proportional change in the t-period spot rate is the same as the proportional change in the one-period spot, or

$$\frac{d(1+S_{0t})}{(1+S_{0t})} = \frac{d(1+S_{01})}{(1+S_{01})}$$

Making this substitution yields

$$\frac{dP_0^t}{P_0^t} = -t\frac{d(1+S_{01})}{(1+S_{01})} \tag{A.1}$$

This equation holds for any t. A coupon-paying bond can be considered a series of pure discount bonds. Let superscripts stand for the time of the flow, and let P_0^t be the current value of the tth-period flow. The price of a bond is the sum of the value of its components. Thus

$$P_0 = P_0^1 + P_0^2 + P_0^3 + \cdots + P_0^T$$

and

$$dP_0 = dP_0^1 + dP_0^2 + dP_0^3 + \cdots + dP_0^T$$

Dividing both sides by P_0 yields

$$\frac{dP_0}{P_0} = \frac{dP_0^1}{P_0} + \frac{dP_0^2}{P_0} + \frac{dP_0^3}{P_0} + \cdots + \frac{dP_0^T}{P_0}$$

or

$$\frac{dP_0}{P_0} = \frac{dP_0^1}{P_0^1}\frac{P_0^1}{P_0} + \frac{dP_0^2}{P_0^2}\frac{P_0^2}{P_0} + \cdots + \frac{dP_0^T}{P_0^T}\frac{P_0^T}{P_0} \tag{A.2}$$

Substituting in the equations for dP_0^1/P_0^1 through dP_0^T/P_0^T and recognizing that P_0^T is the present value of the payment in t yields

$$\begin{aligned}\frac{dP_0}{P_0} &= \frac{\frac{C(1)}{(1+S_{01})}}{P_0}(-1)\frac{d(1+S_{01})}{(1+S_{01})} + \frac{\frac{C(2)}{(1+S_{02})^2}}{P_0}(-2)\frac{d(1+S_{01})}{(1+S_{01})} \\ &\quad + \cdots + \frac{\frac{C(T)}{(1+S_{0T})^T}}{P_0}(-T)\frac{d(1+S_{01})}{(1+S_{01})} \\ &= -\frac{\sum_{t=1}^{T} t\frac{C(t)}{(1+S_{0t})^t}}{P_0}\frac{d(1+S_{01})}{(1+S_{01})} \\ &= -D_2\frac{d(1+S_{01})}{(1+S_{01})}\end{aligned}$$

D_2 measures the sensitivity of bond price to a change in the yield curve where the shift in the yield curve is such that the proportional change in all spot rates is the same.

2. Nonproportional Shift in Spot Rates

D_2 resulted from an assumption that the proportional change in all spot rates is identical. Empirical evidence suggests that long rates change less than short rates. Let $K(t)$ be the proportional change in the tth-period rate compared to the one-period rate. Then

$$\frac{d(1+S_{0t})}{1+S_{0t}} = K(t)\frac{d(1+S_{01})}{1+S_{01}}$$

One way of having long rates less volatile than short rates is if we define $K(t)$ as K^{t-1} and have K less than 1. With this definition the sensitivity of pure discount bonds to a change in interest rates is

$$\frac{dP_0^t}{P_0^t} = -tK^{t-1}\frac{d(1+S_{01})}{1+S_{01}} \tag{A.3}$$

For a coupon bond the proportional change in price is given by Equation (A.2). Substituting (A.3) into (A.2), recalling $P_0^t = C(t)/(1 + S_{0t})^t$, yields

$$\begin{aligned}\frac{dP_0}{P_0} &= \frac{\frac{C(1)}{(1+S_{01})^1}}{P_0}(-1)\frac{d(1+S_{01})}{(1+S_{01})} + \frac{\frac{C(2)}{(1+S_{02})^2}}{P_0}(-2)K^1\frac{d(1+S_{01})}{(1+S_{01})} \\ &\quad + \cdots + \frac{\frac{C(T)}{(1+S_{0T})^T}}{P_0}(-T)K^{T-1}\frac{d(1+S_{01})}{(1+S_{01})}\end{aligned}$$

or

$$\frac{dP_0}{P_0} = -\left[\sum_{t=1}^{T} \frac{tK^{t-1}\dfrac{C(t)}{(1+S_{0t})^t}}{P_0}\right]\frac{d(1+S_{01})}{(1+S_{01})}$$

Define the term in the brackets as D_3. This is the third measure of duration. It measures the sensitivity of bond price to a shift in the yield curve, if the change in the tth-period spot rate is K^t times the change in the one-period spot rate.

Measures of duration have been developed for quite a few other possible changes in the yield curve. For instance, Bierwag has developed a measure of duration for additive changes in the yield curve, multiplicative changes in the yield curve, and the combinations of additive and multiplicative changes. Basically, any reasonable way in which the yield curve can change can give rise to another definition of duration. The problem is that each measure assumes that the yield curve can shift in only one pattern (additive, multiplicative, proportional, etc.) and that once we know the change in one spot rate (e.g., S_{01}), we know the change in all spot rates. In reality, shifts in the yield curve may not follow any set pattern. The true test of the definition will be how effectively it measures the actual changes in the prices of bonds due to a change in the yield curve.

3. Numerical Estimation of Duration

An alternative to estimating duration using the analytical techniques discussed earlier is to estimate it numerically. Table 22.8 shows two hypothesized term structures. The unprimed is the current structure. The primed is a 1% increase in each spot rate in the term structure. From Equation (22.1) and footnote 2, we know that modified or adjusted duration is[17]

$$\frac{dP}{P} = -D_A\, di$$

Thus

$$D_A = \frac{\dfrac{-dP}{P}}{di}$$

Table 22.8 Assumed Term Structures

t	S_{0t}	S_{0t}
1	10	11
2	11	12
3	12	13
4	13	14
5	14	15

[17]Modified duration is normally used because there is often ambiguity in defining $(1 + i)$ when complicated shifts in the term structure are assumed.

We can calculate price assuming both term structures

$$P = \frac{10}{(1.10)} + \frac{10}{(1.11)^2} + \frac{10}{(1.12)^3} + \frac{10}{(1.13)^4} + \frac{110}{(1.14)^5}$$

$$P' = \frac{10}{(1.11)} + \frac{10}{(1.12)^2} + \frac{10}{(1.13)^3} + \frac{10}{(1.14)^4} + \frac{110}{(1.15)^5}$$

$$P = 87.589$$

$$P' = 84.522$$

Thus

$$D = \frac{\dfrac{-(84.522 - 87.589)}{87.589}}{0.01}$$

$$D = 3.5$$

Multiple duration measures can be calculated. For example, it can be assumed that the short and long spot rates can move independent of one another. Movements of all intermediate rates are then linked to movements in these two key rates. A separate duration measure is calculated for movement in each rate. Immunization would be conducted by immunizing against movements in both rates. A possible advantage in numerical estimation is the ability to capture a greater variety of types of shifts in the yield curve. Another advantage is the ability to calculate duration for instruments with call features, because the price determination can reflect the impact of an option being exercised.

4. Duration Measures with Semiannual or Monthly Cash Flows

All of the duration measures were derived in a completely general manner with the length of the period left undefined. However, the reader must be careful in calculating duration for instruments with nonannual coupon payments. The proportional change in Equation (22.1) should be one plus the interest rate for the interval of the interest payments. Furthermore, because it is conventional to express duration in years, the interval should be annualized. For example, consider a bond with 10 years to maturity paying semiannual payments. Then the t in Equation (22.3) would go from 1 to 20 (20 half years) and the resulting duration measure would be cut in half to annualize it. Alternatively, t could be expressed as part of a year or 1/2, 3/2, and so on. In either case, the Δi in Equation (22.1) refers to the change in the six-month rate.

APPENDIX B

EXACT MATCHING PROGRAMS

One of the ways to reduce sensitivity to changes in interest rates is exact matching. Exact matching is an attempt to find the minimum cost portfolio such that the cash flows in each period are sufficient to cover all liabilities. Define the following elements:

1. $L(t)$ as the liabilities in time t
2. $C(t, i)$ as the cash flows in period t from a bond of type i
3. $P(i)$ as the price of bond i
4. $N(i)$ as the number of bond of type i purchased

The cost of the portfolio of bonds is the number of bonds of each type purchased times the price per bond summed over all bonds or $\Sigma_i N(i)P(i)$. This quantity is to be minimized. The aggregate cash flows from all bonds in time t is $\Sigma_i C(t,i)N(i)$. Note that some of these cash flows are coupon payments and some are principal payments. The restriction that cash flows be sufficient to meet liabilities is

$$\sum_i N(i)C(t,i) \geq L(t) \qquad \text{for all } t$$

The final constraint is that the investor cannot issue bonds. This requirement can be stated as $N(i) \geq 0$. Summarizing the exact matching problem is

$$\text{minimize} \sum_i N(i)P(i)$$

subject to

1. $\displaystyle\sum_i N(i)C(t,i) \geq L(t)$ for all t

2. $N(i) \geq 0$ for all i

Note that liabilities are being met by coupon payments or maturing bonds. Bonds are not sold to meet cash flows. Thus the only risk is default risk. Adverse interest rate changes do not affect the ability to meet liabilities. Thus matching programs do not necessitate changes in a portfolio as interest rates change. The foregoing problem is a linear programming problem and can be solved with standard algorithms.

The major variation in this problem is to allow cash carry-forward. If cash can be carried forward, then there are two possible sources of funds that can be used to meet liabilities: cash flows from the bond investment and cash carryover from the prior period.

Let F_t represent the amount of short-term investment and r be the one-period interest rate. Then in time t the value of the short-term investment is the prior period's investment F_{t-1} plus the interest on the investment, or $F_{t-1}(1 + r)$. In each period, sources of funds (cash from the bond portfolio and short-term investments) must be equal to uses of funds (liabilities plus cash to be carried forward):

$$\text{Sources of funds} = \text{Uses of funds}$$

$$\begin{bmatrix}\text{From bond}\\ \text{portfolio}\end{bmatrix} + \begin{bmatrix}\text{From prior short-}\\ \text{term investment}\end{bmatrix} = [\text{Liabilities}] + \begin{bmatrix}\text{New one-period}\\ \text{investment}\end{bmatrix}$$

$$\sum_i N(i)C(t,i) + F_{t-1}(1+r) = L(t) + F_t$$

With the addition of cash carry-forward the problem becomes

$$\text{minimize} \sum_i N(i)P(i)$$

subject to

1. $\displaystyle\sum_i N(i)C(t,i) + F_{t-1}(1+r) \geq L(t) + F_t$ for all t

2. $N(i) \geq 0$ for all i

3. $F_t \geq 0$ for all t

4. $F_{t-1} = 0$

Once again, liabilities are being met out of interest payments and principal payments so that bonds are not being sold. Thus the cash flows from the bond portfolio do not depend on the future course of interest rates. However, r is a future interest rate. If r is set sufficiently low, there will be very little chance that future interest rates will be lower and very little risk that cash flows will be insufficient to meet liabilities. Allowing cash carry-forward cannot result in more cost than not doing so. Thus the formulation allowing cash carry-forward (perhaps at zero interest) provides the better solution. Firms that offer this type of product usually find that competitive pressures force r to approximate current expectations about future short-term rates. In this case the bond matching program becomes much riskier, and once again, its feasibility depends on the actual course of future interest rates.

APPENDIX C

BOND-SWAPPING TECHNIQUES

In the second part of this chapter, we discussed methods of actively managing a bond portfolio. Techniques discussed in these sections allow for bond switches resulting from changes in perceptions of which bonds are over- or underpriced or resulting from changes in risk perceptions. Using the techniques discussed in the second part of this chapter is clearly an appropriate technique for determining bond swaps.

An alternative procedure that makes many fewer assumptions is to attempt to find additional bonds that can be swapped for existing bonds that maintain the future cash flow pattern and yet earn immediate profit from the swap. This is the basic idea underlying a bond swap program. To be specific, define the following elements:

1. $P_B(i)$, the cost of buying bond i
2. $P_S(i)$, the cash received from selling bond i
3. $C(i,t)$, the cash flow of bond i in period t
4. $N_B(i)$, the number of bonds of type i purchased
5. $N_S(i)$, the number of bonds of type i sold

With these definitions the cost of the bonds purchased is

$$\sum_i N_B(i)P_B(i)$$

The profit is the difference between the proceeds from the sale and the cost of the purchase, or

$$\sum_i N_S(i)P_S(i) - \sum_i N_B(i)P_B(i)$$

The object of a bond swap program is to maximize this difference subject to not reducing cash flows. If the swap does not result in reduced cash flows, the bond portfolio will still meet any liabilities.

To express this constraint, we write

$$\sum_i N_B(i)C(i,t) \geq \sum_i N_S(i)C(i,t) \qquad \text{for all } t$$

One swap model is

$$\text{maximize } \sum_i N_S(i)P_S(i) - \sum_i N_B(i)P_B(i)$$

subject to

1. $\sum_i N_B(i)C(i,t) \geq \sum_i N_S(i)C(i,t)$ for all t
2. $N_B(i), N_S(i) \geq 0$ for all i

The ability to carry forward funds from an earlier to a later period can be added to the bond swap problem. This increases the risk because future interest rates are unknown. However, it increases the number of swap opportunities. Adding the ability to carry forward funds can be developed as follows. Let

1. F_t be the short-term investment in period t
2. r be the one-period interest rate

The value of the cash carried forward from period $t - 1$ is $F_{t-1}(1 + r)$. The investment in short-term cash in period t is F_t.

If short-term borrowing is not allowed, then F_t must be nonzero. The complete problem is

$$\text{maximize} \sum_i N_S(i)P_S(i) - \sum_i N_B(i)P_B(i)$$

subject to

1. $\sum_i N_B(i)C(i,t) + F_{t-1}(1+r) \geq \sum_i N_S(i)C(i,t) + F_t$ for all $t = 1, \ldots, T$
2. $N_B(i), N_S(i) \geq 0$ for all i
3. $F_t \geq 0$ for all t
4. $F_{t-1} = 0$

This is the standard bond swap problem.

APPENDIX D

CONVEXITY

In this appendix we derive Equation (22.4) and show how the mathematical definition of convexity is derived.

The formula for the first three terms in a Taylor series expansion of a function $f(i + h)$ in the region of i as h approaches zero:

$$f(i+h) = f(i) + \frac{f'(i)(h)}{1} + \frac{f''(i)(h)^2}{2\times 1} + \cdots$$

where the prime denotes derivatives.

Define $P(i)$ as the price of a bond at an interest rate i. Then, writing the price of the bond at a new interest rate $(i + h)$ using the series expansion results in

$$P(i+h) = P(i) + P'(i)h + 1/2\, P''(i)h^2 \tag{D.1}$$

The price of the bond is

$$P(i) = \sum_{t=1}^{T} \frac{C(t)}{(1+i)^t}$$

Then the first derivative with respect to $(1 + i)$ is

$$P'(i) = \sum_{t=1}^{T} \frac{tC(t)}{(1+i)^t} \frac{1}{1+i} \tag{D.2}$$

and the second derivative is

$$P''(i) = \sum_{t=1}^{T} \frac{t(t+1)C(t)}{(1+i)^t} \frac{1}{(1+i)^2} \tag{D.3}$$

The return due to a change in interest rates is

$$R_i^{\mu} = \frac{P(i+h) - P(i)}{P(i)}$$

And using Equations (D.1), (D.2), and (D.3) together with the fact that Δ_i as defined in the text is $h/(1 + i)$,

$$R_i^{\mu} = -D_i \Delta_i + C_i (\Delta_i)^2$$

where

$$D_i = \left[\sum_{t=1}^{T} \frac{tC(t)}{(1+i)^t} \right] \Big/ P(i)$$

$$C_i = 1/2 \left[\sum_{t=1}^{T} \frac{t(t+1)C(t)}{(1+i)^t} \right] \Big/ P(i)$$

Alternative convexity measures could be derived by allowing the discount rate to vary over time.

QUESTIONS AND PROBLEMS

1. Consider a bond with semiannual coupon payments of $50, a principal payment of $1,000 in 5 years, and a price of $1,000. Assume that the yield curve is a flat 10%. What is the duration of the bond?

2. Consider a bond with annual coupon payments of $100, a principal payment of $1,000 in 10 years, and a cost of $1,000. Assume a flat yield curve with a 10% yield to maturity. What is the duration of the bond? If the yield curve remains unchanged, what is the bond's duration in three years? In five years? In eight years?

3. Given the following bonds:

Bond	Duration (years)
A	5
B	10
C	12

construct three different portfolios of the three bonds, each with a duration of nine years.

4. Assume liabilities of $250, $500, and $550 must be met in periods 1, 2, and 3, respectively. Find a portfolio of the bonds shown below that meets these cash outflows. What is the cost of the portfolio? (*Hint:* The question does not require a least-cost portfolio.

Thus the linear programming procedure of Appendix B is not necessary.)

		Cash Flows in Period		
Bond	Price	1	2	3
A	950	50	1,050	
B	1,000	100	100	1,100
C	920	1,000		

5. Assume that the yield curve for the data of Problem 3 is 10%. Further assume that the three bonds are of equal value and the only bonds existing. Set up a single-index representation of their covariance. What is the covariance between all pairs of bonds?

BIBLIOGRAPHY

1. Attari, Mukarram. "Discontinuous Interest Rate Processes: An Equilibrium Model for Bond Option Prices," *Journal of Financial and Quantitative Analysis*, **34,** No. 3 (Sept. 1999), pp. 293–322.
2. Babble, D. "Duration and the Term Structure of Interest Rate Volatility," in G. Bierwag, G. Kaufman, and A. Toevs (eds.), *Innovations in Bond Portfolio Management: Duration Analysis and Immunization* (Greenwich, CT: JAI Press, 1983).
3. Bierwag, G. O. "Immunization, Duration and the Term Structure of Interest Rates," *Journal of Financial and Quantitative Analysis*, **12,** No. 4 (Dec. 1977), pp. 725–742.
4. Bierwag, G. O., and Kaufman, George. "Coping with the Risk of Interest Rate Fluctuations: A Note," *Journal of Business*, **50,** No. 3 (July 1977), pp. 364–370.
5. ——. "Immunization Strategies for Funding Multiple Liabilities," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1983), pp. 113–124.
6. ——. "Durations of Non-Default-Free Securities," *Financial Analyst Journal*, **44,** No. 4 (July/Aug. 1988), pp. 39–46.
7. Bierwag, G. O., Kaufman, G. G., and Khang, C. "Duration and Bond Portfolio Analysis: An Overview," *Journal of Financial and Quantitative Analysis*, **13,** No. 4 (Nov. 1978), pp. 671–681.
8. Bierwag, G. O., Kaufman, George G., and Toevs, Alden. "Single Factor Duration Models in a Discrete General Equilibrium Framework," *Journal of Finance*, **37,** No. 2 (May 1982), pp. 325–338.
9. ——. *Innovations in Bond Portfolio Management: Duration Analysis and Immunization* (Greenwich, CT: JAI Press, 1983).
10. Boardman, Calvin M., and McEnally, Richard W. "Factors Affecting Seasoned Corporate Bond Prices," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 2 (June 1981), pp. 193–206.
11. Boquist, J. A., Racette, G. A., and Schlarbaum, G. "Duration and Risk Assessment for Bonds and Common Stocks," *Journal of Finance*, **30,** No. 5 (1975), pp. 1360–1365.
12. Brennan, M. J., and Schwartz, E. "Savings Bonds, Retractable Bonds and Callable Bonds," *Journal of Financial Economics*, **5** (1977), pp. 67–88.
13. ——. "A Continuous Time Approach to the Pricing of Bonds," *Journal of Banking and Finance*, **3** (1979), pp. 133–155.
14. ——. "Conditional Predictions of Bond Prices and Returns," *Journal of Finance*, **35** (1980), pp. 405–417.
15. ——. "Duration, Bond Pricing, and Portfolio Management," in G. Bierwag, G. Kaufman, and A. Toevs (eds.), *Innovations in Bond Portfolio Management: Duration Analysis and Immunization* (Greenwich, CT: JAI Press, 1983).
16. Brown, Stephen, and Dybvig, Philip. "The Empirical Implications of the Cox, Ingersoll Ross Theory of the Term Structure of Interest Rates," *Journal of Finance*, **41,** No. 3 (1986), pp. 617–632.

17. Campbell, John Y. "Who Should Buy Long-Term Bonds?" *American Economic Review*, **91,** No. 1 (March 2001), pp. 99–127.
18. Carr, J. L., Halpern, P. J., and McCallum, J. S. "Correcting the Yield Curve: A Re-interpretation of the Duration Problem," *Journal of Finance*, **29,** No. 4 (1974), pp. 1287–1294.
19. Chambers, Donald R., Carleton, Willard T., and McEnally, Richard W. "Immunizing Default-Free Bond Portfolios with a Duration Vector," *Journal of Financial and Quantitative Analysis*, **23,** No. 1 (March 1988), pp. 89–104.
20. Constantinides, George M., and Ingersoll, Jonathan E., Jr. "Optimal Bond Trading with Personal Taxes," *Journal of Financial Economics*, **13,** No. 3 (Sept. 1984), pp. 299–351.
21. Cornell, Bradford, and Green, Kevin. "The Investment Performance of Low-Grade Bond Funds," *Journal of Finance*, **46,** No. 1 (March 1991), pp. 29–48.
22. Cox, J. C., Ingersoll, J. E., Jr., and Ross, S. A. "Duration and the Measurement of Basic Risk," *Journal of Business*, **52,** No. 1 (Jan. 1979), pp. 51–61.
23. Dothan, U. L. "On the Term Structure of Interest Rates," *Journal of Financial Economics*, **6** (1978), pp. 59–69.
24. Ehrhardt, Michael C. "A New Linear Programming Approach to Bond Portfolio Management: A Comment," *Journal of Financial and Quantitative Analysis*, **24,** No. 4 (Dec. 1989), pp. 533–537.
25. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Fundamental Economic Variables, Expected Returns, and Bond Fund Performance," *Journal of Finance*, **50,** No. 4 (Sept. 1995), pp. 1229–1256.
26. Elton, Edwin J., Gruber, Martin, J., and Michaely, Roni. "The Structure of Spot Rates and Immunization," *Journal of Finance*, **XLV,** No. 2 (June 1990), pp. 621–641.
27. Elton, E., Gruber, M., and Naber, P. "Bond Returns, Immunization and the Return Generating Process," in M. Sarnat and G. Szego (eds.), *Studies in Banking and Finance, Essays in Memory of Irwin Friend* (New York: North-Holland, 1988).
28. Fisher, L., and Weil, R. L. "Coping with the Risk of Interest Rate Fluctuations: Returns to Bondholders from Naive and Optimal Strategies," *Journal of Business*, **44,** No. 3 (Oct. 1971), pp. 408–431.
29. Fong, Gifford H., and Vasicek, Oldrich A. "The Tradeoff between Return and Risk in Immunized Portfolios," *Financial Analyst Journal*, **34,** No. 5 (Sept./Oct. 1983), pp. 73–78.
30. ——. "A Risk Minimizing Strategy for Portfolio Immunization," *Journal of Finance*, **39,** No. 5 (Dec. 1986), pp. 1541–1546.
31. Hessel, Christopher A., and Huffman, Lucy. "The Effect of Taxation on Immunization Rules and Duration Estimation," *Journal of Finance*, **36,** No. 5 (Dec. 1981), pp. 1127–1142.
32. Ingersoll, J. "Is Immunization Feasible?" in G. Bierwag, G. Kaufman, and A. Toevs (eds.), *Innovations in Bond Portfolio Management: Duration Analysis and Immunization* (Greenwich, CT: JAI Press, 1983).
33. Ingersoll, J. E., Jr., Skelton, J., and Weil, R. L. "Duration Forty Years Later," *Journal of Financial and Quantitative Analysis*, **13** (Nov. 1978), pp. 627–650.
34. Liebowitz, Martin L., and Weinberger, Alfred. "Contingent Immunization Part II: Problem Areas," *Financial Analysts Journal*, **39,** No. 1 (Jan./Feb. 1983), pp. 35–50.
35. Livingston, M., and Caks, J. "A 'Duration' Fallacy," *Journal of Finance*, **32** (March 1977), pp. 185–187.
36. Macaulay, F. R. *Some Theoretical Problems Suggested by the Movements of Interest Rates, Bond Yields, and Stock Prices in the United States since 1865* (New York: Columbia University Press, 1938).
37. Marshall, William J., and Yawitz, Jess B. "Lower Bounds on Portfolio Performance: An Extension of the Immunization Strategy," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1982), pp. 101–114.
38. Niederhoffer, V., and Regan, P. "Earnings Changes, Analysts' Forecasts, and Stock Prices," *Financial Analysts Journal*, **28,** No. 3 (May–June 1972), pp. 65–71.
39. Nelson, J., and Schaefer, S. "The Dynamics of the Term Structure and Alternative Portfolio Immunization Strategies," in G. Bierwag, G. Kaufman, and A. Toevs (eds.), *Innovations in Bond Portfolio Management: Duration Analysis and Immunization* (Greenwich, CT: JAI Press, 1983).

40. Prisman, Eliezer Z. "Immunization as a Maximum Strategy," *Journal of Business Finance*, **20,** No. 4 (Dec. 1986), pp. 491–509.
41. Redington, F. M. "Review of the Principles of Life-Office Valuations," *Journal of the Institute of Actuaries*, **18** (1952), pp. 286–315.
42. Richard, S. F. "An Arbitrage Model of the Term Structure of Interest Rates," *Journal of Financial Economics*, **6** (1978), pp. 33–57.
43. Richards, Malcolm. "Analysts' Performance and the Accuracy of Corporate Earnings Forecasts," *Journal of Business*, **49,** No. 3 (July 1976), pp. 350–357.
44. Ronn, Ehud I. "A New Linear Programming Approach to Bond Portfolio Management," *Journal of Financial and Quantitative Analysis*, **22,** No. 4 (Dec. 1987), pp. 439–466.
45. Schaefer, Stephen M., and Schwartz, Eduardo S. "A Two-Factor Model of the Term Structure: An Approximate Analytical Solution," *Journal of Financial and Quantitative Analysis*, **19,** No. 4 (Dec. 1984), pp. 413–421.
46. Vasicek, P. "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics* (Nov. 1977), pp. 177–188.

23

Option Pricing Theory

The markets for options are among the fastest-growing markets for financial assets in the United States. While option trading is not new, it experienced a gigantic growth with the creation of the Chicago Board of Options Exchange in 1973. The listing of options meant more orderly and thicker markets for these securities.

The growth in option trading has been accompanied by a tremendous interest among academics and practitioners in the valuing of option contracts. In this chapter we discuss alternative types of options, examine the effect of certain characteristics on the value of options, and present explicit models for valuing options.

TYPES OF OPTIONS

An option is a contract entitling the holder to buy or sell a designated security at or within a certain period of time at a particular price. There are a large number of types of option contracts, but they all have one element in common: the value of an option is directly dependent on the value of some underlying security. Options represent a claim against the underlying security and thus are often called contingent claim contracts. The two least complex options are called puts and calls. These are the most widely traded options. In addition, most other options either can be valued as combinations of puts and calls or can be valued by the methodology developed to value puts and calls. Consequently, we begin this section with a discussion of puts and calls, and then we discuss other types of options and combinations of basic options.

Calls

The most common type of option is a call. A call gives the owner the right to buy a fixed number of shares of a stock at a fixed price, either before or at some fixed date. It is common to refer to calls that can be exercised at only a particular point in time as European calls and calls that can be exercised at any time up to, and including, the expiration date as American calls. Take, for example, a November 20 American call on Mobil at $70. This call gives the owner the right to buy a certain number of shares of Mobil at $70 a share anytime on or before November 20. Calls are normally traded in units of 100 shares. Thus

one call would be a right to buy 100 shares of Mobil. Each characteristic of the call has a name. For example, the $70 price is called the exercise price. The final date at which the call can be exercised is the expiration date.[1]

One of the distinguishing characteristics of a call is that if it is exercised, the exchange of stocks is between two investors. One investor issues the call (termed the *call writer*) and the other investor purchases the call. The call is a side bet between two investors on the future course of the security. Figure 23.1*a* shows the profit per share of stock for the holder of a call at the expiration date. The figure represents the pattern for a call originally purchased for $5 with an exercise price of $50. For a stock price below $50, it would not pay to exercise the call because shares could be purchased in the open market for less than the exercise price. For share prices above $50, it would pay to exercise the call and gain by the difference between the share price and exercise price. For example, if the share price is $54, then the holder of a call benefits from the ability to purchase the stock at $50 rather than $54. For share prices up to $55, the owner of the call loses money since the payoff from the stock purchase is less than the cost of the call. For a stock price above $55, there is a profit.

The position of the call writer is depicted in Figure 23.1*b*. The pattern of the profit is exactly opposite that of the call purchaser. For a stock price below $50, the call writer makes a profit equal to the $5 per share received from the issuance of the call; from $50 to $55, part of the $5 is lost by having to furnish the stock at a price below the market price; above $55, the call writer loses more than was received by selling the call.

Up to now, we have referred to shares being traded between individuals as a result of exercise at the expiration of a call. We could have also discussed the exercise of a call before the expiration date. However, we have not done this so far because calls (even those that have an exercise price below the price of the stock) are rarely exercised before the

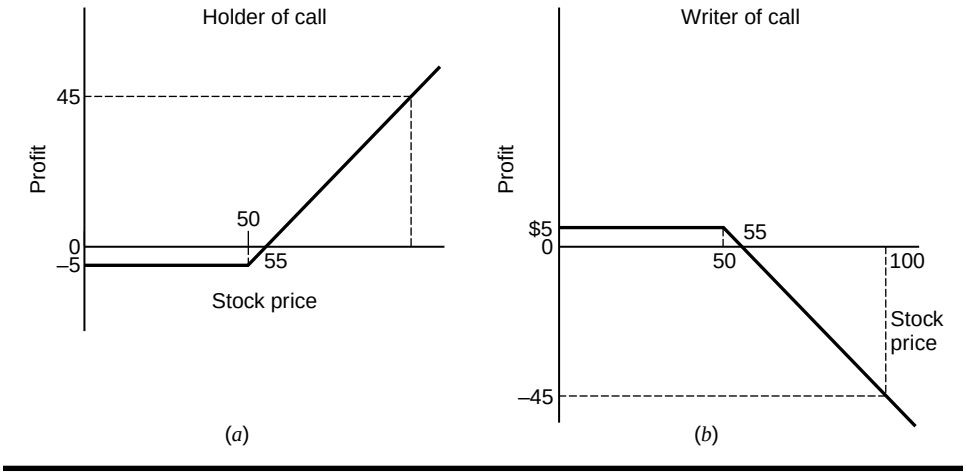


Figure 23.1 Profit from call.

[1]In the example, we use an arbitrary date for expiration. For options not listed on the exchanges, any date is possible. However, options trading on the Chicago Board of Options have standardized expiration dates. Any single security will normally have options outstanding with three different expiration dates. These dates are three months apart (e.g., April, July, and October). These options expire at 10:59 A.M. Central Time on the Saturday after the third Friday of the month.

expiration date.[2] For example, assume the share price is $60 and the exercise price is $50. Clearly, a profit can be made by exercising the option. There is a third alternative. Instead of exercising the option, sell it.

The sale may be to someone who does not currently maintain a position in the option or it may be to an investor who wrote an option and also wishes to liquidate his position.[3] The listing of options on exchanges facilitates these sales. With options listed on the exchange, the mechanics of the purchase or writing of an option becomes identical to the mechanics of the purchase or sale of a stock, except for differences in margin requirements.

There are actions that a firm might take that will affect the value of its shares. For example, a two-for-one stock split would be expected to cut the price of a share in half. Stock dividends and cash dividends are two other examples. The value of an option is affected by these actions of the firm. Clearly, if there were no adjustment in the exercise price when a stock splits, the value of an option would be substantially reduced. Most options are protected against stock dividends and stock splits by automatic adjustments in the exercise price and the number of shares that can be purchased with one option. Cash dividends are not as frequently protected against. For example, there are no adjustments for cash dividends for options traded on the exchanges. The price of a stock on average decreases by slightly less than the amount of the dividend when a stock goes ex-dividend. Thus, all other things being equal, the price of an option should be lower on a stock that will go ex-dividend before the expiration date.

The next most common type of option is a put, which we will discuss in the next section.

Puts

A put is an option to sell stock at a given price on or before a particular expiration date. Consider, for example, a $50 General Electric put of December 18. The person who owned such a put would have the right to sell the General Electric stock to the person who issued the put at $50 a share on or before December 18. Puts, like calls, are traded in units of 100 shares. Thus one put involves the right to sell 100 shares. If the exercise can take place only at the expiration date, it is called a European put. If the exercise can take place at any time on or before the expiration date, it is called an American put. A put, like a call, involves a transaction between two investors. Thus the writing of puts has no effect on the value of the firm.

Figure 23.2 shows the profit at the expiration date for a put with an exercise price of $50 that originally cost $5. Figure 23.2*a* shows the profit to the owner of the put. Figure 23.2*b* shows the profit to the writer of a put. Consider Figure 23.2*a*. For prices above $50, the owner of the put would prefer to sell shares in the regular market rather than to the writer of the put, because the price received is greater. Thus, for prices above $50, the exercise value is zero. For prices above $45 but below $50, the owner of the put would prefer to exercise her option instead of selling her stock on the open market. However, the owner of the put loses money, because she paid more for the put than she gains from the sale at a

[2]In a later section of this chapter we will discuss the well-established proof that (except for possible exceptions associated with dividend payments) it never pays to exercise an American call prior to the expiration date. It is always better to sell it rather than exercise it.

[3]When an individual sells an option, the person purchasing it need not be the original writer. Rather, the individual purchasing the option is whoever happens to wish to buy the option on the day of sale. This is identical to what happens with any other security. When you buy a share of stock and subsequently sell it, the individual from whom you buy or to whom you sell is unknown and normally different.

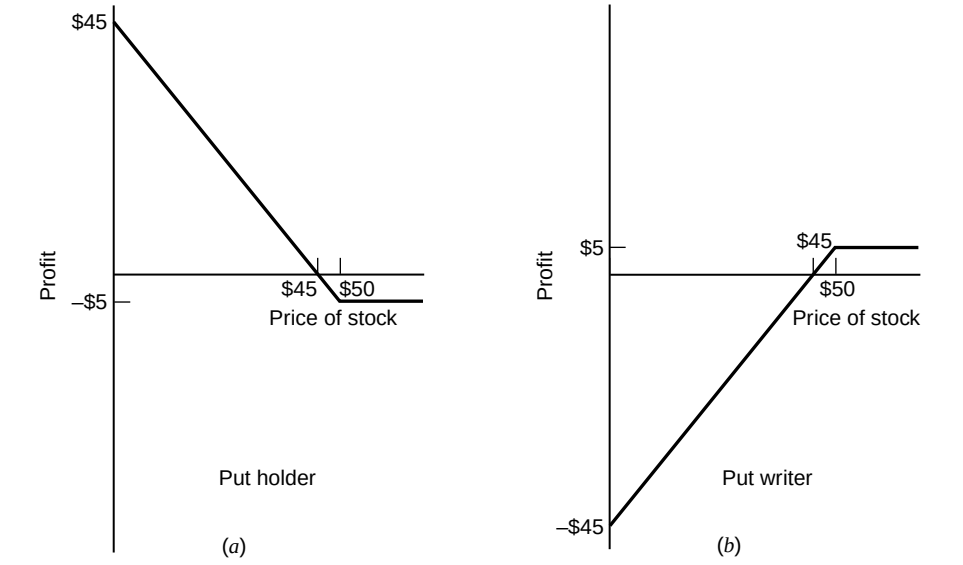


Figure 23.2 Profit from put.

higher price. Below $45, the owner of the put makes money, because the amount she gains from the sale at a more attractive price more than compensates for the cost of the put. The payoff pattern for the writer of the put is the exact opposite of the payoff pattern for the owner. For prices above $45, he makes money, and for prices below $45, he loses.

Puts, like calls, are rarely exercised before expiration. Assume the share price in our example declined to $40. At $40, it clearly pays to exercise rather than to let the option expire. Instead of exercising the option, the owner could sell the right. Although an American put is more likely to be exercised before expiration than an American call, we will show that it generally pays to sell rather than exercise a put, for the sale price will almost always be higher than the exercise value. The exception can occur when the put is deep in the money.

Warrants

A warrant is almost identical to a call. Like a call, it involves the right to purchase stock at an exercise price at or before an expiration date. A warrant differs from a call in one way: a warrant is issued by the corporation rather than another investor. This seemingly small difference is very important. There are two instances when this difference has an effect on the value of the firm that issues the warrant. First, when the warrants are issued, the company receives the money for the warrant. Second, when the warrants are exercised, the following occurs:

1. The company receives the exercise price.
2. The number of shares of the firm that are outstanding goes up by the number of shares that are exercised.
3. The number of warrants still outstanding goes down.

Calls and puts are side bets by market investors, and the corporation has no direct interest in transactions involving these options, either when they are created or when they are

exercised. Warrants, on the other hand, are used by the corporation to raise capital. The corporation and its shareholders have a definite interest in their issuance and exercise, because these transactions affect both the amount of cash the firm has raised and the ownership interest of its shareholders. Because the issuance and exercise of warrants affect the value of the security on which the warrant represents a contingent claim, the valuation of warrants becomes a more complex problem than the valuation of calls.

Combinations

Part of the fun of reading the options literature is the colorful terminology. One of the areas where it is especially colorful is the naming of combinations of options. An infinite number of combinations of puts and calls can be considered. A combination of a put and call with the same exercise price and expiration date is called a *straddle*. A similar combination of two puts and a call is a *strip*. If the combination is two calls and a put, it is called a *strap*. The payoff pattern at expiration is easy to determine using the techniques discussed earlier. Similarly, the valuation can be accomplished using the techniques discussed later in this chapter.

Consider a straddle. Figure 23.3*a* shows the profit at expiration from the point of view of the purchaser of the option. Figure 23.3*b* is the profit from the point of view of the writer. As we can see from examining these diagrams, a straddle should be purchased by someone who believes the price of the shares will move substantially either up or down, without being sure of the direction, and who also believes that other investors have underestimated the magnitude of future price changes. For example, a straddle could be purchased by someone who knew that major information was about to be announced that would seriously affect the company's fortune, was unsure whether the information would be good news or bad news, and believed that other investors were unaware of the existence of this information. In contrast, the writer of a straddle is an investor who believes that the share price will trade at close to the exercise price, while others believe differently.

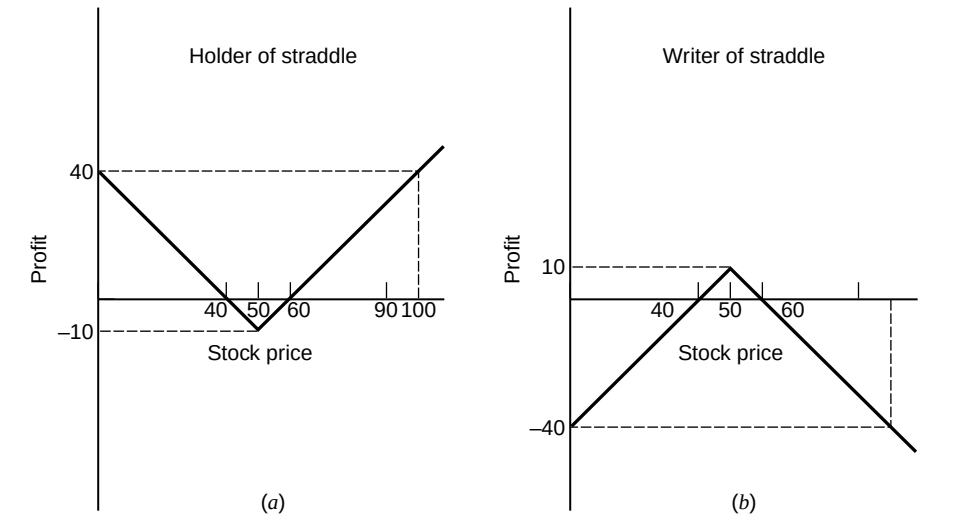


Figure 23.3 Profit from straddle.

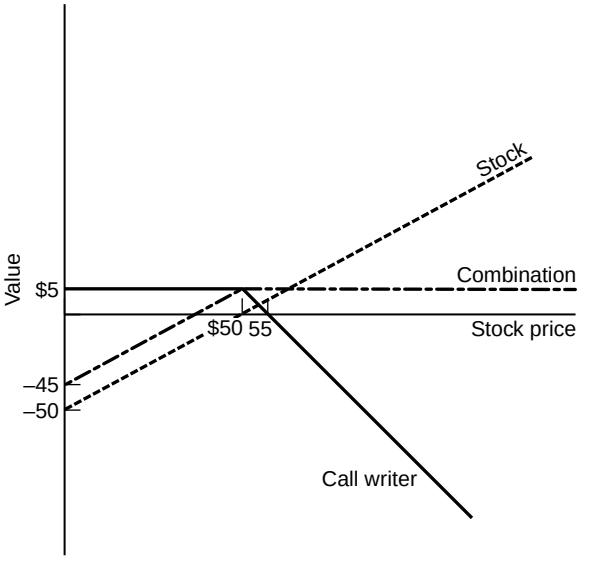


Figure 23.4 The value of a combination of common stock and a call.

One of the interesting ways to trade options is in combination with the stock on which they represent a claim. The investor who combines stocks with contingent claims on the stocks has two assets with very strong correlations. Consider the writer of a call who also owns the shares.[4] Figure 23.4 shows the payoff pattern at expiration. The exercise price is assumed to be $50, the cost of the stock to the holder is also assumed to be $50, and the call is assumed to cost $5. Three separate lines are shown: one for the stock, one for the call, and one for the combination. As Figure 23.4 shows, an investor who writes a call and owns the stock rather than simply owning the stock increases the return at low stock prices at the expense of returns at the higher share prices.

As a final example, consider the ownership of a put plus the ownership of stock. Once again, assume an exercise price of $50, a stock cost of $50, and a put cost of $5. Figure 23.5 shows the payoff pattern. This combination reduces the return at higher stock prices in exchange for guaranteeing that if the stock declines in price, the portfolio will not decline below a lower limit.

Another type of combination is an option that can be purchased only in combination with another security. A convertible bond is an example of this combination. A convertible bond has the same characteristics as a normal bond and in addition can be converted into the shares of a company. Thus the convertible bond can be considered a bond plus a call. However, the call has a special feature. Conversion of the bond into shares of stock involves giving up the bond, plus sometimes cash for the stock. Because the value of the bond changes over time, the exercise price changes over time.

We have discussed a number of combinations of options and options plus security positions in this section. Many others are possible. We leave it to our readers to determine the payoff pattern for those they find interesting.

[4]The writing of a call while owning the stock is called *writing a covered call.*

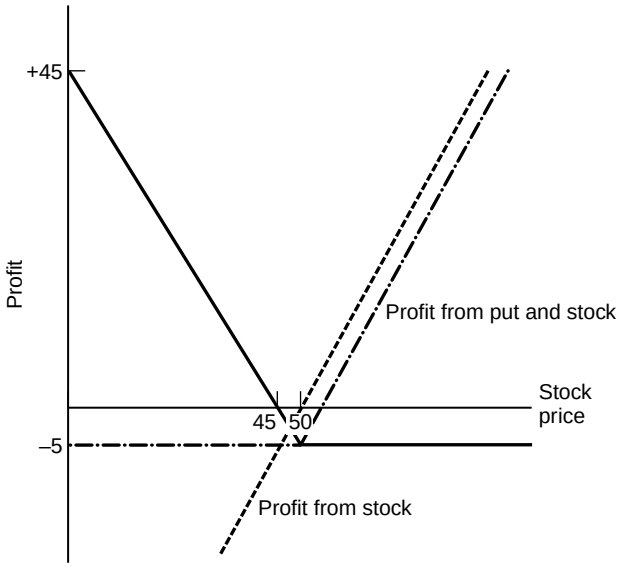


Figure 23.5 Profit from put and stock.

SOME BASIC CHARACTERISTICS OF OPTION VALUES[5]

In a short time we will examine formal option valuation models. However, before we do so, we can infer the manner in which certain characteristics of options should affect their value in a rational market. Not only are these relationships interesting in themselves, but they will also prove useful as a check on valuation models developed in the later sections. Any valuation model should be consistent with these basic relationships. It is interesting to note that some of the earlier option valuation models that were later proved incorrect were not consistent with these basic relationships.

Relative Prices of Calls with Alternative Characteristics

Recall that the European call gives the holder the right to purchase stock at the exercise price on a particular date (the expiration date). The American call differs from the European call in that it can be exercised at any time up to the expiration date. Because the American call is a European call with the added opportunity to exercise before the expiration date, it cannot be worth less than the European call. Thus the first relationship established is that a European call with the same expiration date and exercise price as an American call cannot sell for more than the American call.

Consider two American calls with the same exercise prices and assume both calls are on the same stock. The one with the longer life offers the investor all the exercise opportunities of the one with the shorter life, plus some additional opportunities. Hence it cannot be worth less. It might bother the reader that we do not simply say that the longer-lived call is more valuable. In general, this is true, but in some extreme cases (e.g., when both calls are worthless), this is not true. Hence the more cautious statement.

[5]The results in this section were developed by Merton (1973).

The next relationship concerns the exercise price. Consider two calls with the same expiration date written on the same stock. The one with the higher exercise price cannot be more valuable than the one with the lower exercise price. This is obvious, because the holder of the latter can be in the same position as the holder of the former, upon exercise, except that she will have cash left over.

While these relationships seem quite simple, as discussed earlier, not all valuation models that have been developed were consistent with these principles; hence they are worth keeping in mind.

Minimum Value of a European Call

In this section we will show that the value of a European call on a non-dividend-paying stock is at least the greater of zero and the difference between the stock price and the present value of the exercise price. To see this, consider two different portfolios. Portfolio *A* involves the purchase of a call and a bond that matures at the expiration date of the call and which at that date will have a value equal to the exercise price. If *R* is the interest rate between the time the call is valued and the expiration date, and if *E* is the exercise price, then bonds in the amount of $E/(1 + R)$ should be purchased. An alternative to portfolio *A* is the purchase of stock directly. Call this portfolio *B*. The key characteristics to these investments are shown in Table 23.1. S_1 is the stock price at expiration, S_0 is the current stock price, *E* is the exercise price, and *C* is the current price of the call. The payoffs at the expiration date are shown in the last two columns of the table.

If $S_1 > E$, then the payoffs from both portfolios are the same. However, if $S_1 \leq E$, the payoff from portfolio *A* is larger. Thus portfolio *A* is at least as desirable as portfolio *B*, and if $S_1 \leq E$ at expiration is possible, *A* is more desirable. Given that portfolio *A* is at least as desirable as *B*, it cannot cost less than *B*; otherwise no one would purchase the stock (portfolio *B*). Therefore

$$C + \frac{E}{1+R} \geq S_0$$

or

$$C \geq S_0 - \frac{E}{1+R}$$

The European call cannot sell for less than the stock price less the present value of the exercise price. Because the call cannot sell for a price below zero, we have completed the proof.

Table 23.1 Payoffs from Alternative Holdings

		Value at Expiration Date	
Action	Investment	If $S_1 > E$	If $S_1 \leq E$
Portfolio *A*			
Buy call	$-C$	$S_1 - E$	0
Buy bonds	$\frac{-E}{1+R}$	E	E
Total	$-C - \frac{E}{1+R}$	S_1	E
Portfolio *B*			
Buy stock	$-S_0$	S_1	S_1

Early Exercise of an American Call

Probably the most surprising conclusion of modern option pricing theory is that it never pays to exercise an American call before the expiration date on a stock that does not pay dividends or whose exercise price is adjusted for dividend payments. Later we will present a simple proof. But before we do, it is worthwhile to discuss why this holds. The reason is simple but subtle. The American call is worth more alive than dead. It is worth more keeping the American call alive by not exercising it than killing it through exercise. Thus an investor no longer wishing to hold the call is better off selling it than exercising it. Consider an example. Assume a stock is selling for \$60 and an investor holds an American call with an exercise price of \$50. Furthermore, assume this investor believes that the stock price will decline between now and the expiration date. Clearly the investor would prefer to exercise the call now rather than hold it and exercise it at a later date. There is another option: sell the call to another investor.

If the call has a market price higher than the \$10 the investor makes on exercise (\$60 stock price − \$50 exercise price), selling the call is preferable. Why should the price of the call be more than \$10? The American call has two sources of value: the value of an immediate call (\$10) plus the value of the chance to call from now to the expiration date. As long as this latter opportunity has value, the American call should sell for more than \$10. You might well ask why someone would wish to buy the call when the investor believes the stock price will decline. The answer is that this cannot be the general market belief or the stock price would have already declined. In other words, the aggregate market belief must be that the correct price is \$60 and that at \$60, the total return from the stock is competitive with securities of similar risk. Thus the market must believe the return on the stock will be positive.

Now for the proof. Earlier, we argued that an American call cannot be worth less than a European call. We also showed that the European call was worth more than the maximum of zero and the difference between the stock price and the present value of the exercise price $[S_0 - E/(1 + R)]$. Thus the value of the American call must be greater than the maximum of zero and $S_0 - E/(1 + R)$. However, if the call is exercised, its value is $S_0 - E$. Because $S_0 - E/(1 + R) > S_0 - E$, the call sells for more than its value if exercised.

The foregoing discussion assumed that the stock did not pay a dividend before the expiration date or that the call was protected against dividends by having the exercise price adjusted by the amount of the dividend. If the stock is dividend paying or the call is not protected, early exercise is possible. Consider the example discussed earlier with a \$60 stock price and a \$50 exercise price. If the stock was about to pay a large dividend, then investors could rationally believe that the share price should be lower than \$60 between the ex-dividend date and the expiration date and thus that the current difference is the best that can be obtained.[6]

Put Call Parity

A put and the underlying stock can be combined in such a way that the combination has the same payoff pattern as a call. Similarly, a call and the underlying equity can be combined so that they have the same payoff pattern as a put. This allows the put or call to be priced in terms of the other security.

[6]Stock prices are expected to drop by slightly less than the amount of the dividend when a stock goes ex-dividend.

This relationship is easiest to derive for European options. Furthermore, it is convenient to assume that the common equity will not pay a dividend in the period before the option expires. Define

S_0 as the current stock price

S_1 as the stock price at the expiration date

E as the exercise price

C as the call price

P as the put price

R_B as the borrowing rate

R_L as the lending rate

Now consider a combination of a share of stock, a put, and taking a loan for an amount $E/(1 + R_B)$. If $E/(1 + R_B)$ is borrowed and if the interest rate between the purchase of the combination and the expiration date is R_B, then $[E/(1 + R_B)]\ (1 + R_B) = E$ will have to be paid back. Thus, if $E/(1 + R_B)$ is borrowed, an amount equal to the exercise price will have to be paid back at the expiration date. The payoff of this combination at the expiration date is shown in Table 23.2. The payoff pattern is, of course, exactly the same pattern as for a call.

The investor has two possible investments: the call or the portfolio being discussed. Each investment has the same value at the expiration date. If they sell at different prices currently, then the investor can purchase the least expensive investment and issue the more expensive investment. Because they have the same payoff pattern at the expiration date, the investor can use the proceeds of the one investment to meet the obligations of the other. If they have different costs, a guaranteed profit can be made. Assuming the portfolio is less expensive than the call, then the investor would write the call and purchase the portfolio. If the call is more expensive than the portfolio, this combination then yields a guaranteed profit. The guaranteed profit is immediate and has zero risk. Such a possibility cannot last long in any efficiently functioning market. Thus the call cannot be more expensive than the portfolio, and writing the call plus purchasing the portfolio cannot be profitable. Writing a call involves a cash inflow of C and purchasing portfolio A involves flows of $-S_0 - P + E/(1 + R_B)$. This implies

$$C - S_0 - P + \frac{E}{1 + R_B} \leq 0$$

or

$$S_0 + P - \frac{E}{1 + R_B} \geq C$$

Table 23.2 Payoffs of Portfolios Involving Puts

	Value at Expiration Date	
Security	If $S_1 > E$	If $S_1 \leq E$
Portfolio A		
Buy stock	S_1	S_1
Buy put	0	$E - S_1$
Borrow	$-E$	$-E$
Total	$S_1 - E$	0
Purchase of call		
Buy call	$S_1 - E$	0

Consider what happens if the call is less expensive than the portfolio. In this case, the investor would wish to issue the portfolio and buy the call. The flows would be $-C$ for the call and $S_0 + P - E/(1 + R_L)$ for portfolio A.

These flows closely resemble those discussed earlier, but R_L has replaced R_B. Because we assume the investor is short selling the portfolio rather than purchasing it, the investor is lending rather than borrowing, and R_L is assumed to be the lending rate.

If the call is less expensive than the portfolio, then this combination yields a guaranteed profit. A guaranteed profit with no risk can't last long in the market, so buying the call and issuing the portfolio cannot be a profitable combination. This implies that

$$S_0 + P - \frac{E}{1+R_L} - C \le 0$$

or

$$C \ge S_0 + P - E/(1+R_L)$$

Putting the equations together yields

$$S_0 + P - \frac{E}{1+R_B} \ge C \ge S_0 + P - \frac{E}{1+R_L}$$

If $R_L = R_B$, the preceding inequalities become equalities, and we have the put call parity relationship.

Some comment on the two different arbitrage combinations is in order. The first combination was appropriate if the call was more expensive than the portfolio. This strategy involved buying stock and a put, borrowing, and writing a call. All of these are feasible, and the combination is a full description of the necessary actions.[7] The other combination was appropriate when the call was less expensive than the portfolio. This involved selling the stock short, writing a put, lending, and buying a call. The analysis assumed that the proceeds of the short sale were immediately available. This is unrealistic in general, as discussed earlier. However, it would represent a realistic situation for an investor who currently owned the shares and who engages in a transaction identical to a short sale by selling his existing shares. Because there are likely to be many of these investors, the put call parity theorem should hold reasonably well.

The previous analysis examined the payoff pattern at the expiration date of the option. This is, of course, the only relevant date to examine for European options. With American options, other dates are potentially relevant. One of the components of the portfolio is a put. It can be shown that it may pay to exercise a put before expiration, and the value of the American put may be higher than shown in the prior tables.[8] The issuance of an American put involves the risk of premature exercise, and the arbitrage discussed earlier need no longer hold. Another problem with applying the prior analysis is the possibility of the payment of dividends. The payment of dividends would, of course, affect the payoffs depicted earlier. If the dividends are already announced, then the stock price can be adjusted by reducing it by the present value of the dividends. With this adjustment, dividends do not affect the prior analysis, except insofar as they affect the probability of exercising a put. If dividends are not announced, then adjusting by the expected dividends is

[7]The only margin required is the margin on the call. The ownership of the stock is sufficient to meet this requirement.

[8]See Merton (1973).

reasonably satisfactory. All these issues mean that the put call parity relationship may not hold perfectly for American options. Nevertheless, it should be a close approximation to market relationships. This is exactly what the empirical results (see Klemkosky and Resnick, 1979; Gould and Galai, 1974) have shown.[9]

VALUATION MODELS

In this section we will present and discuss two widely used option valuation models. The models we will present are for the European call. From the last section the reader will recall that it never pays to exercise an American call before its expiration date if it is either dividend protected or the stock will not pay dividends before the expiration date. An American call that meets these conditions will not be exercised before it expires, and thus it can be valued as a European call. In the previous section we derived the relationship between the value of puts and calls. Thus the valuation formula for a call can also be used to value puts.

The differences in modern valuation formulas stem from the alternative assumptions made about how share price changes over time. In this section we will present two models. One assumes that the percentage change in share price follows a binomial distribution; the other assumes it follows a lognormal distribution.

Binomial Option Pricing Formula

The simplest of the option pricing formulas is the binomial option pricing formula.[10] Because the implications of the formula are similar to those of more complicated formulas and because the formula is easy to derive and understand, we will present a detailed derivation in this section.

Assume that a call is being valued one period before expiration. Further assume that the stock is currently selling at \$50 and will either increase to \$75 or decrease to \$25. Further assume that the borrowing and lending rate is 25%. Under these conditions, what is the current value of a call with an exercise price of \$50?

To answer this question, consider the portfolio shown in Table 23.3.

The way the portfolio is constructed, the investor receives nothing at period 1, whether the stock sells at \$25 or \$75. This suggests that the investment should cost nothing or that $2C - 50 + 20 = 0$ or that $C = \$15$. To confirm this, consider two other values of C: $C = \$10$ or $C = \$20$. If $C = \$10$, then the call is underpriced. This suggests buying the call,

Table 23.3 Cash Flows on a Zero-Payoff Portfolio

		Flows at 1	
	Flows at 0	$S_1 = 25$	$S_1 = 75$
Write 2 calls	$+2C$	0	-50
Buy 1 share of stock	-50	$+25$	$+75$
Borrow \$20	$+20$	-25	-25
		0	0

[9]The arbitrage involving the short sale of stock is sometimes profitable empirically. This part of the put call relationship has less empirical support.

[10]The earliest derivation of this formula is in Stone (1969). Sharpe (1978), Cox and Ross (1976), and Rendleman and Bartter (1979) have independently derived the formula.

Table 23.4 Cash Flows on a Zero-Payoff Portfolio

		Flows at 1	
	Flows at 0	$S_1 = 25$	$S_1 = 75$
Purchase 2 calls	−20	0	+50
Short 1 share of stock	+50	−25	−75
Lend $20	−20	+25	+25
	+10	0	0

shorting the stock, and lending will lead to an instantaneous profit. Let us examine this combination in Table 23.4.

No matter which share price occurs at period 1, there are no net flows. The only flow occurs at zero and is a plus $10. This is a guaranteed return with no risk, and as investors purchase the combination of securities shown before, prices will adjust until the profit disappears.

Now consider the case $C = \$20$. At this price, the call is overpriced and the investor issues the call, borrows, and buys stock. The flows are shown in Table 23.5.

Once again, there are no net flows at period 1, so that if this situation existed, the investor would have a guaranteed return with no risk. Such opportunities should disappear quickly if they exist, and the three securities should be so priced that riskless profits cannot occur. The call must sell at $15. Let us generalize this example.

The portfolio was constructed so that payoffs from the call plus the stock were the same, no matter what the value of the stock at time period 1. Then, by lending or borrowing, the payoff of the portfolio of calls, stock, and riskless bonds can be made to have zero return at time 1. In the example given, a combination of two calls and one share of stock yielded 25—no matter what happened at period 1—and served the purpose. The number of shares of stock per call that makes the payoff from the combination independent of share price is called the hedge ratio. Let

S_0 = the stock price at period zero

E = the exercise price of the option

u = 1 plus the percentage change in stock price from time 0 to time 1, if the stock price increases

d = 1 plus the percentage change in stock price from time 0 to time 1 if the stock price decreases

C = the call price

α = the number of shares of stock purchased per share of the call

C_u = the value of the call if the stock increases in value (the maximum of $uS_0 - E$ or 0)

C_d = the value of the call if the stock decreases in value (the maximum of $dS_0 - E$ or 0)

Table 23.5 Cash Flows on a Zero-Payoff Portfolio

		Flows at 1	
	Flows at 0	$S_1 = 25$	$S_1 = 75$
Write 2 calls	+40	0	−50
Buy 1 share of stock	−50	+25	75
Borrow $20	+20	−25	−25
	+10	0	0

Table 23.6 Cash Flows from a Portfolio of Calls and Stock

Action	Flows at 0	Flows at 1	
		$S_1 = uS_0$	$S_1 = dS_0$
Write call	C	$-C_u$	$-C_d$
Buy α shares of stock	$-\alpha S_0$	$\alpha u S_0$	$\alpha\, dS_0$

Consider Table 23.6. For this to be a hedged portfolio, the flows at period 1 must be independent of the value of the stock. Thus

$$-C_u + \alpha u S_0 = -C_d + \alpha\, dS_0$$

or

$$\alpha = \frac{C_u - C_d}{S_0(u-d)}$$

In the previous example, $C_d = 0$, $C_u = 25$, $S_0 = 50$, $u = 1.5$, and $d = 0.5$. Thus

$$\alpha = \frac{25-0}{50(1.5-0.5)} = \frac{25}{50} = \frac{1}{2}$$

Thus, to have the call plus the stock have the same payoffs, no matter what value the stock has at period 1, we must purchase one-half as many shares of stock as we write calls. Two calls and one share of stock, the hedged position used in the previous example, is consistent with this ratio. Utilizing a hedge ratio of α means that the flows at time 1 are the same or $-C_u + \alpha u S_0 = -C_d + \alpha d S_0$. To make the portfolio flows at one equal zero, we borrow an amount such that we owe $(C_d - \alpha d S_0)$ at time 1 (or, equivalently, $C_u - \alpha u S_0$). If r is 1 plus the interest rate, we borrow $(C_d - \alpha d S_0)/r$. This results in the flows shown in Table 23.7.

As discussed earlier, if the flows at period 1 on the portfolio are zero, the investment also must be zero. Thus

$$C - \alpha S_0 - \frac{C_d - \alpha d S_0}{r} = 0 \qquad \text{or} \qquad C = \frac{\alpha r S_0 + C_d - \alpha d S_0}{r} \tag{23.1}$$

Substituting for α yields

$$C = \frac{\left(\frac{(C_u - C_d)}{S_0(u-d)}\right) r S_0 + C_d - \left(\frac{C_u - C_d}{S_0(u-d)}\right) d S_0}{r}$$

or

$$C = \frac{\frac{(C_u - C_d)r + C_d(u-d) - d(C_u - C_d)}{u-d}}{r}$$

or

$$C = \frac{C_u \frac{(r-d)}{(u-d)} + C_d \frac{(u-r)}{(u-d)}}{r}$$

Table 23.7 Cash Flows on a Zero-Payoff Portfolio of Stock and Calls

		Flows at 1	
Action	Flows at 0	Price = uS	Price = dS
Write call	C	$-C_u$	$-C_d$
Buy α stock	$-\alpha S_0$	$\alpha u S_0$	$\alpha d S_0$
Borrow	$\dfrac{-C_d + \alpha d S_0}{r}$	$C_d - \alpha d S_0 = C_u - \alpha u S_0$	$C_d - \alpha d S_0$
Total	$C\ \alpha S_0 - \dfrac{C_d - \alpha d S_0}{r}$	0	0

This is the formula for the value of the call with one period remaining until it expires. It can be further simplified by defining $P = (r - d)/(u - d)$. With this definition,

$$(1-P) = 1 - \frac{r-d}{u-d} = \frac{(u-d)-(r-d)}{u-d} = \frac{u-r}{u-d}$$

Making these substitutions into the previous formula, we have

$$C = \frac{C_u P + C_d(1-P)}{r}$$

where

$$P = \frac{r-d}{u-d}$$

Before proceeding, one comment is in order. Notice that in this derivation we were never concerned with the probability of an up or down movement. We have never even discussed what it might be. P and $1 - P$ are not probabilities; rather, they are numbers that depend on the magnitude of the up and down movements and the riskless rate of interest. What does the value of the call depend on? Examining the formula shows that it depends on C_u, C_d, r, u, and d. However, C_u and C_d depend on the exercise price, the size of u and d, and the current stock price S_0. For example, if an up movement in the stock involves an exercise, then $C_u = uS_0 - E$. Thus, in a two-period example, the call price ultimately depends on

u, the size of the up movement

d, the size of the down movement

E, the exercise price

r, one plus the riskless rate of interest

S_0, the current stock price

The type of factors that affect the call price carry over to the more complicated model discussed later.

There is a second way this formula can be derived that yields useful insight into the valuation of options. If we use the value of α derived earlier as the ratio of stocks to calls, then no matter whether the stock goes up or down, we get the same return. An investment that has the same outcome no matter what happens is riskless and should yield the riskless

rate of interest. Thus, if we buy the stocks while writing sufficient calls to maintain the hedged position given by α, the return on the investment must be r:

$$\text{(investment) } r = \text{outcome}^{11}$$

$$(\alpha S_0 - C)r = \alpha\, dS_0 - C_d$$

A glance at Equation (23.1) shows that it is identical to the preceding expression. To move from Equation (23.1) to the option pricing formula involved substituting for α and rearranging. Thus both procedures led to the same result. The idea of valuing options by forming a riskless hedge carries over to models of more complicated stock movements that will be examined in a later section of this chapter.

The formula for pricing a call when there is more than one period to the expiration is a simple extension of the one-period formula just derived. Figure 23.6 shows what can happen to the share price when there are two periods to go to expiration.

The formula just derived allows us to determine the value of the call with one period to expiration (e.g., at period 1). However, knowing the value at time 1 allows the calculation of the value at time 0 by acting as if there is one period to go. In this iterative manner the binomial valuation can be derived. In Appendix A at the end of the chapter we go through a detailed derivation and show that the value of the call with n periods to go is

$$C = S_0 B[a, n, P'] - Er^{-n} B[a, n, P]$$

where

$$P = \frac{r-d}{u-d} \quad P' = \frac{u}{r} P$$

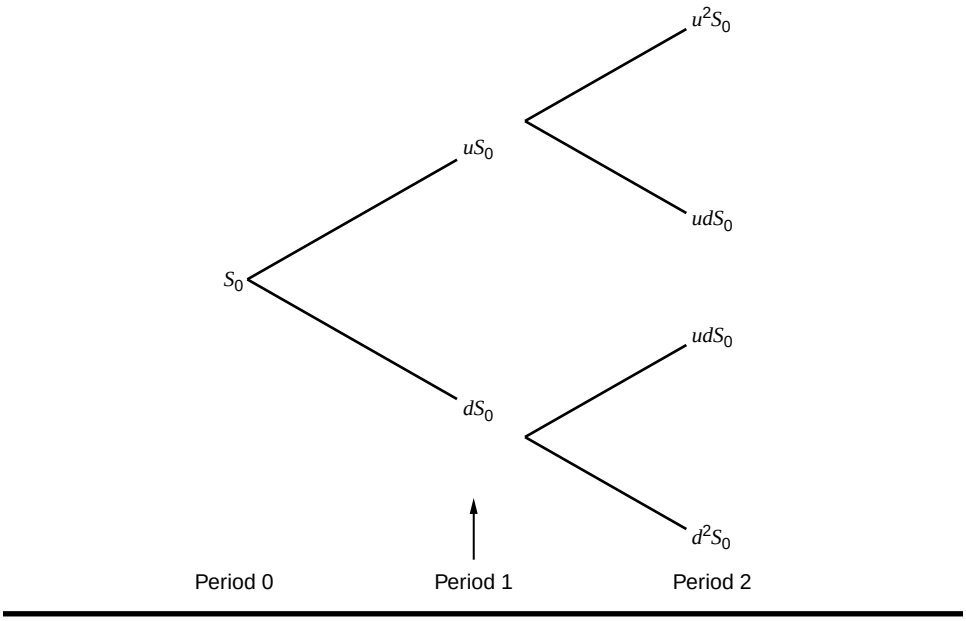


Figure 23.6 The movement of stock prices through time.

[11]The outcome could alternatively have been written as $\alpha u S_0 - C_u$.

and

S_0	is the current stock price
E	is the exercise price
n	is the number of periods to expiration
r	is 1 plus the riskless rate of interest
a	is the lowest number of upward moves in price at which the call takes on a positive value at expiration
$B[a,n,P']$	is the probability of a number of up moves in share price equal to or greater than a occurring out of n movements where the probability of an up move is P' (the probability is obtained from the binomial formula or can be looked up in a table of the binomial formula)
u and d	remain as defined earlier

Some additional comment on $B[a,n,P']$ or $B[a,n,P]$ is warranted. First, a is determined by examining the current price, the exercise price, and the expiration date. Assume, for example, that the current stock price was \$50, the exercise price was \$60, $u = 1.50$, $d = 0.80$, and $n = 10$. A little calculation will show that if there are four or more increases in share price, the stock price will exceed \$60 by the expiration date. Thus, $a = 4$ in this example.[12] The second comment necessary is that although in order to calculate $B[a,n,P']$ or $B[a,n,P]$ we act as if P' or P are probabilities, in actuality they have nothing to do with probabilities. P and P' depend on the size of the up and down movements and the risk-free rate. They are not connected with the probabilities of these up and down movements taking place. We refer to them as probabilities solely because we employ them as if they were probabilities in using the binomial formula.

The reader might well wonder how to determine the inputs in the binomial option pricing formula. In particular, how are the up and down price movements (u and d) determined? Values of u and d are set so that the return distribution resulting from their values is what the user considers reasonable. In practice the user specifies the standard deviation of the stock and the number of intervals until expiration over which a movement up or down takes place, and then calculates a value of u and d that would result in the return process having the standard deviation that was specified. The specific formulas are

$$u = e^{+\sigma\sqrt{t/n}}$$
$$d = e^{-\sigma\sqrt{t/n}}$$

where

n	is the number of intervals until expiration
σ	is the annual continuous time standard deviation of the return on the stock (the standard deviation of the log of returns)
t	is the time to expiration in years
e	is the exponential function

Specifying a larger number of intervals increases the number of possible returns the stock can have over the period but increases the computational burden. A fairly small number of intervals seems to produce accurate option valuation.

[12] $50(1.50)^4(0.80)^6 = 66.35$, whereas $50(1.50)^3(0.80)^7 = 35.39$. More formally, a can be defined as the number for which $u^{a-1}d^{n-(a-1)}S_0 < E \leq u^a d^{n-a}S_0$.

Note that each of the factors that we demonstrated as affecting call price in the two-period model also affects call prices in the multiperiod model. The call is a function of the size of the up movement, the size of the down movement, the exercise price, the current share price, and the riskless rate of interest. In addition, the multiperiod model is a function of *n*, the number of periods remaining until expiration.

The binomial formula just derived can be utilized to derive two other valuation formulas that allow a continuous change in the share price. This is accomplished by letting the length of the period between up or down movements become very small, and hence the number of periods is very large. The most popular of these models is due to Black and Scholes and is developed in the next section.

The Black–Scholes Option Valuation Formula

In the previous section of this chapter we derived an option pricing formula under the assumption that the rate of return on the underlying stock followed a binomial formula. As the number of time periods gets very large, the binomial distribution converges to the normal distribution.

If we assume that a stock's continuously compounded rate of return follows a normal distribution, then the option pricing model developed in the preceding section reduces to the Black–Scholes option pricing formula presented below:[13]

$$C = S_0 N(d_1) - \frac{E}{e^{rt}} N(d_2) \tag{23.1a}$$

$$d_1 = \frac{\ln(S_0/E) + \left(r + \frac{1}{2}\sigma^2\right)t}{\sigma\sqrt{t}} \tag{23.1b}$$

$$d_2 = \frac{\ln(S_0/E) + \left(r - \frac{1}{2}\sigma^2\right)t}{\sigma\sqrt{t}} \tag{23.1c}$$

where

r = the continuously compounded riskless rate of interest
C = the current value of the option
S_0 = the current price of the stock
E = the exercise price of the option
e = 2.7183
t = the time remaining before the expiration date expressed as a fraction of a year
σ = the standard deviation of the continuously compounded annual rate of return
$\ln(S_0/E)$ = natural logarithm of S_0/E
$N(d)$ = the value of the cumulative normal distribution evaluated at d

The Black–Scholes formula can be used to value any option. In the next section of this chapter we will discuss how to use it. Before we do, we will discuss the variables that affect the valuation of calls as well as the relationship of this formula to that discussed earlier.

[13]See Cox and Ross (1976).

Perhaps the most interesting aspect of the Black–Scholes model is a variable that does *not* appear as a determinant of the value of a call. This variable is the expected rate of return on the stock. Any of the option models determines the price of the option in terms of the price of the underlying stock. The stock price, in fact, acts as the numeraire in which call prices are expressed. Expected return enters the model insofar as it determines current share price, but given current share price, it does not affect the value of the call.

The impact of the other variables on the value of the call can be seen by examining the properties of the Black–Scholes model as each changes. In general, the results are as follows: the higher the ratio of the current price of the stock to the exercise price of the call, the higher the value of the call. This is reasonable, for the higher this ratio, the less the price of the stock must increase for the call to have a value on its expiration date. The longer the time to maturity on the call, the higher the value of the call. This again is sensible, for the longer the time to maturity, the more the stock's price is likely to deviate from its present level at maturity. Since the payoff from deviations from price is asymmetrical, the longer time to maturity increases the value of the call.[14] Finally, the higher the riskless rate of interest, the greater the value of the call. This follows logically from the fact that the higher the riskless rate, the lower the present value of the amount that must be paid to exercise the call. The reader should note that these conclusions are consistent with the general statements we said must hold in a rational option pricing formula. They are also consistent with the conclusions we derived when we discussed the binomial formula.

Using the Black–Scholes Model In examining the Black–Scholes formula we saw that the only data we needed to value an option were the current price of the stock, the exercise price of the option, the time remaining before expiration of the option, a cumulative normal probability table, the riskless rate of interest, and the standard deviation of the continuously compounded annual rate of return on the stock. All of these, except for the standard deviation, are easily observable.[15] One way to estimate the standard deviation of the continuously compounded annual rate of return on a stock is to use historic data on stock returns.[16] The Black–Scholes model was derived under the assumption of identically distributed rates of return over time. If this assumption in fact were strictly true over all periods, then estimates of the variance from historical data would be very good. As an example of this procedure, assume that we wish to estimate the appropriate variance for some stock using one year of historical weekly data. The price relative for each stock is simply the price at the end of the week plus any dividends divided by the price at the beginning of the week. The natural logarithm of the price relative is the continuously compounded rate of return per week. The standard deviation of the continuously compounded rate of return can easily be computed by applying the standard formula to the sequence of continuously compounded rates of return. For example, standard deviation is

$$= \left(\sum_{i=1}^{N} \left(\frac{\left(X_i - \overline{X}\right)^2}{N} \right) \right)^{1/2}$$

[14]For example, if the price of the stock is below the exercise price, then decreases in price up to the exercise time would result in the same value, zero, at the exercise time. In contrast, a rise in price could lead to a positive value for the call at the exercise time.

[15]The continuously compounded riskless rate of interest is usually found by taking the rate on a government security that has a maturity date equal to (or as close as possible to) the expiration date on the call.

[16]In the next section of this chapter we will discuss another method that uses the Black–Scholes model itself to prepare estimates of the standard deviation.

To convert the continuously compounded weekly standard deviation to a yearly standard deviation, simply multiply by the square root of 52.

The Black–Scholes model assumes that interest rates are continuously compounded. Because interest rates are generally stated using discrete compounding, some calculations are required to convert to the continuously compounded rate.

Assume the risk-free rate is calculated as ending value of the bond minus beginning value divided by beginning value (a discrete rate). As an example, assume the calculation results in a risk-free rate of 6%. Then the continuously compounded rate used in the Black–Scholes formula is r in the following formula:

$$e^{-r} = 1.06 \qquad \text{or} \qquad r = 0.0582$$

Once inputs for the Black–Scholes valuation formula have been defined, one can easily solve for the value of a call option. Perhaps this can best be illustrated with an example:

$$\begin{aligned} S_0 &= 90 \\ E &= 100 \\ t &= 0.5 \text{ (6 months)} \\ \sigma &= 0.5 \\ r &= 0.10 \end{aligned}$$

Then d_1 and d_2 can be easily computed as follows:

$$d_1 = \frac{\ln(90/100) + \left(0.10 + \frac{1}{2}(0.25)\right)(.5)}{0.5\sqrt{0.5}} \approx 0.02$$

$$d_2 = \frac{\ln(90/100) + \left(0.10 - \frac{1}{2}(0.25)\right)(.5)}{0.5\sqrt{0.5}} \approx -0.33$$

From any table of the cumulative normal distribution, we can compute

$$\begin{aligned} N(d_1) &= N(0.02) = 0.5080 \\ N(d_2) &= N(-0.33) = 0.3707 \end{aligned}$$

The value of the call is

$$C = 90(0.5080) - \frac{100}{e^{.10(0.5)}}(0.3707) = \$10.46$$

Using the Black–Scholes formula, we now have a theoretical value for the call of \$10.46. Assume that the call was selling at \$9.50. If the Black–Scholes formula is correct, the call is undervalued in the market. The investor can take advantage of this by buying the call directly. Alternatively, the investor could be protected against adverse stock price changes by buying the call and selling the stock short. Recall from the previous section that this combination is a riskless hedge.[17] It can be shown that if we accept the Black–Scholes option pricing formula as correct, the appropriate hedge ratio is given by $N(d_1)$ or, in our example, 0.5080. This means that for every call option purchased, 0.5080, or slightly more than half of the share of stock, should be sold short.

[17]In the section discussing the binomial formula we derived a hedge ratio. A similar argument in the Black–Scholes model shows that the hedge ratio is $N(d_1)$. Examining d_1 shows that it should be expected to change over time, and thus the hedge ratio also changes.

The hedge ratio is sufficiently important to traders that it has been given a name of its own. It is called *delta*. The construction of portfolios to take advantage of any mispricing in either options or underlying securities is known as delta hedging. While at any moment in time the hedge ratio can easily be determined from the formulas for *d* and *N(d)* presented earlier, examination of the formulas makes it clear that delta hedging is not a passive activity. This is because the size of the hedge changes with a change in the price of the underlying security, the passage of time, or a change in the volatility of the underlying security. The rate of change in the hedge ratio with respect to a change in the price of the underlying asset is known as *gamma*. The rate of change in the price of the option with respect to time is called *theta*. The rate of change of the option with respect to the volatility of the underlying asset is known as *vega*. Obviously, the smaller the size of gamma, theta, or vega, the less often hedge ratios have to be adjusted and the easier and less costly it is to maintain a hedge portfolio. These parameters of hedging are sufficiently important so they are routinely computed by traders in the option market, and the exact formula for computing them can be found in any advanced text on options (e.g., Hull, 2001).

Many traded calls are on securities that pay dividends over the life of the option. From previous discussion, early exercise, if it occurs, will occur immediately before the stock goes ex-dividend. Thus, a call will be optimally exercised either at maturity or just before the ex-dividend date. This pattern, that early exercise only occurs just before the ex-dividend date, can be used to value a call. The investor can view the problem of valuing a call on a dividend-paying stock as owning two calls—one that expires just before the ex-dividend date and one that expires at maturity. The value of the actual call is very close to the maximum value of each call considered separately. The call expiring just before the ex-dividend date is valued by the standard formula with the time to expiration, taken as the time between the current date, and the day before the ex-dividend date. The other call is valued similarly with two changes. First, the time to expiration is whatever it is for the call. Second, the price used in the option formula is the current price less the present value of the dividend. The logic is that when the stock goes ex-dividend, the value of the stock drops by the amount of the dividend. Unlike a stockholder, the option holder does not receive the dividend, so the current value of the stock to the option holder is reduced by the dividend. Because the dividend is paid in the future, the current loss is the present value of the dividend.

Let us consider an example. Assume the following:

$S_0 = 50$
$E = 50$
$r = 3\%$ for 90 days
$r = 2\%$ for 60 days
$D = \$5$
$\sigma = 0.20$

Further assume the stock goes ex-dividend in 61 days and matures in 90 days. Then the value of the option, assuming exercise is 60 days just before it goes ex-dividend, is found by assuming $S_0 = 50$, $E = 50$, $r = 2\%$, and $\sigma = 0.20$. The option value using the Black–Scholes formula is $1.68.

Similarly, the value of the option, assuming exercise at maturity, is found by assuming $E = 50$, $r = 3\%$, $\sigma = 0.20$, and that S_0 is

$$S_0 = 50 - 5e^{-(0.02)(60/360)} = 45.02$$

The exponential *e* simply finds the present value of the dividend when continuous compounding is used.

The Black–Scholes option value, assuming no early exercise, is \$0.41. The maximum of these two calculations is \$1.68, and this would be considered the minimum option value. The call option value would be greater than these two numbers because of the opportunity to reconsider the decision immediately before the ex-dividend date.

Implicit Estimates of Stock's Own Variance from Option Formulas In the previous section of this chapter we discussed the input needed to use option valuation models. All of the model input variables were easily observed, except for one—the variance of the instantaneous rate of return on the stock. Up to now, we have assumed that the value of this variable is inferred from historical data. However, there is a second way in which option valuation formulas such as the Black–Scholes formula can be used. If we believe that option prices are such that the Black–Scholes model holds on average, then the market price of the option can be substituted for C in the model. The only remaining unknown in the formula is the instantaneous rate of variance of the stock.[18] Because we have one equation and one unknown, a formula like the Black–Scholes formula can be used to determine the variance of the stock. If the assumptions behind the Black–Scholes model are completely valid, and the model holds on average, then the variance implied by the Black–Scholes model should be a good estimate of the market's expectation about the variance of a stock's return. On any one stock, there are likely to be many calls outstanding, and these calls will probably have different exercise prices and expiration dates. From each of these calls we can obtain an estimate of the standard deviation of the stock's continuously compounded rate of return. The efficiency of the estimate should be improved if we combine several independent estimates. Ways of doing this will now be discussed.

The simplest way to find an estimate of σ is to take an average of the estimates obtained from each call outstanding on the stock. If there are N calls outstanding, and if σ_j is the estimate of the standard deviation arrived at by employing data for the jth call, then

$$\sigma = \frac{1}{N}\sum_{j=1}^{N}\sigma_j$$

Not all authors weight the estimates equally. Many authors place less weight on estimates obtained from calls that have prices less sensitive to σ. This weighting scheme would place less weight on calls where the stock price is far from the exercise price and more weight on calls where the stock price and exercise price are close. There are several variants of this weighting.[19] Some authors simply discard estimates from calls where the stock price is very different from the exercise price. Other authors have suggested weighting by the relative sensitivity of the call price of the option to changes in the standard deviation.[20]

---

[18]The Black–Scholes formula cannot be explicitly solved for variance. However, an iterative procedure can be used to find the implied variance for any stock that is consistent with this formula. See Latane and Rendleman (1976) for a discussion of search procedures.

[19]See Boyle (1977) and Schmalensee and Trippi (1978) for additional suggestions as to plausible weighting schemes.

[20]This technique was used by Latane and Rendleman (1976). Defining $\partial C_j/\partial\sigma_j$ as the change in the call price of call j to a change in standard deviation of call j, then the weight on the jth estimate of standard deviation (W_j) is

$$W_j = \frac{\partial C_j}{\partial \sigma_j}\Bigg/\sum_k \frac{\partial C_k}{\partial \sigma_k}$$

and

$$\sigma = \sum_{j=1}^{N} w_j \sigma_j$$

Although calculating weights in the manner just discussed is only one of a large number of weighting techniques that have been advocated for arriving at estimates of the variance of a stock's return, it is one of the few subject to empirical tests. Latane and Rendleman (1976) have used a weighting scheme similar to that described earlier to investigate the ability of estimates of variance from the Black–Scholes model to serve as forecasts of the future. To judge the usefulness of this technique, Latane and Rendleman perform two sets of tests. One set looks directly at whether better forecasts of actual future variance are achieved when (1) forecasts are prepared by computing variance over a historical period or (2) forecasts are prepared from the Black–Scholes model. They conclude that forecasts from the Black–Scholes model are more accurate. As a second test, they examine whether large profits are earned by arbitraging mispriced calls where the value of the call is computed by using the variances arrived at in (1) or (2). They again conclude that the use of variances inferred from the Black–Scholes model leads to a better valuation of assets (a higher excess return) than does the forecasting of variances from historical data. For an excellent analysis of forecasting variance, see Figlewski (1997).

While this technique for estimating variance has important implications for the pricing of options, it also can be important for portfolio selection. In earlier chapters, we discussed how estimates of expected returns, variances, and correlation coefficients are necessary inputs to the portfolio selection process. We devoted two chapters to estimating correlation coefficients. We also mentioned that estimates of expected returns must come from security analysts and that analysts can be trained to produce estimates of variances. The latter is much more difficult than the former. The option literature seems to provide either a useful alternative measure of variances or, at the least, a useful benchmark to help the analysts in their estimation process.

ARTIFICIAL OR HOMEMADE OPTIONS

One of the existing insights in modern option theory is that an appropriate mixture of Treasury bills and a security creates a payoff pattern identical to the pattern of an option on the underlying security. This is exciting because options are written only on a limited number of securities, and artificially created options can produce the payoff pattern of an option on securities or portfolios where actual options do not exist. Consider, for example, an arbitrary portfolio. Assume further that the portfolio does not resemble an index. In this case options would not exist on the portfolio. Assume further the portfolio has a value of $100. A homemade put at $105 can be created. This eliminates the risk of returns below 5% for the portfolio. Of course, homemade puts, like traded puts, have a cost. In the case of an artificial put, the cost comes in the form of a reduction in returns when returns on the portfolio are above 5%. Thus a homemade put changes the return distribution of the portfolio by eliminating returns below 5% and reducing the returns above 5% in the same manner as a traded put would. Whether this is desirable depends on the investor's taste for risk and return. The creation of an artificial put on a portfolio goes by the name of "portfolio insurance" because the portfolio is insured against returns below 5%.

Let us examine in more detail the construction of an artificial put. The first row of Table 23.8 shows the payoff pattern of a put if the stock price can end up at $50 or $40 in one period and if the exercise price is $45. Rows 2 and 3 show a combination of shorting the stock and buying T-bills that has the same payoff pattern as the put. If the put does not exist, then shorting the stock and owning T-bills creates a homemade put that has the same payoff pattern as a publicly traded put.

Note that when a homemade put is written in conjunction with a portfolio or asset, the investor is not literally short. Consider a pension portfolio and an artificial put. The short

Table 23.8 Illustration of Homemade Put

	Value at Expiration Date	
	If Stock Price Is 40	If Stock Price Is 50
Buy put	5.00	0.00
Short $\frac{1}{2}$ share of stock	−20.00	−25.00
Buy T-bills	25.00	25.00
Sum	5.00	0.00

sale is accomplished by selling off part of the portfolio. Holding less of the portfolio is equivalent to owning the portfolio while simultaneously being short part of it.

Note also that the homemade put is created by selling off less than one share of stock. In the example it was one-half share. If there were more than one period, the fraction of shares sold short would change over time. Thus the creation of homemade options involves frequent readjustment of the combination of the underlying security and T-bills in order that the payoff pattern resembles that of an option. Early implementation of this idea involved literally selling and buying shares of an asset or portfolio. To replicate the payoff pattern of an option, frequent transactions were called for. Because frequent sales and purchases meant substantial transaction costs, shares were traded less frequently, and the payoff pattern deviated substantially lower transaction costs. Furthermore, features can be used to construct an asset like stocks or bonds. Thus the growth of futures markets has been a spur to the creation of artificial options.

USES OF OPTIONS

In earlier sections of this chapter, we discussed the nature of options and their valuation. In this section, we will examine the major uses of options by individual investors and institutional investors.

Modifying the Return Pattern

In Chapter 5 we discussed the efficient frontier with riskless lending and borrowing. The efficient frontier was a straight line such as that shown in Figure 23.7. Note that as we increase the number of Treasury bills in the portfolio, we lower expected return and the standard deviation of return. However, we do not fundamentally change the distribution of return. If we plot two portfolios such as A, which is half Treasury bills and half risky assets, and portfolio B, which is 100% in risky assets, to examine the distribution of returns, we get the distributions shown in Figure 23.8. Note that the effect of reducing risk by adding Treasury bills is to squeeze the distribution and shift it to a lower mean return. Adding Treasury bills does not fundamentally change the shape of the return distribution. If we assume that riskless lending and borrowing was not possible, then the same conclusion holds. Moving along the efficient frontier changes the mean return and variance but does not change the shape of the distribution.

One of the major uses of options is to modify the shape of the return distribution. Consider an index fund with the index currently valued at $1,500. Furthermore, assume the investor believes the expected return is 10% with a standard deviation of 15%. Then, the investor holding the index fund expects to face a probability distribution such as that shown as a solid curve in Figure 23.9. Assume the investor buys a put on the S&P index with an exercise price of $1,550. Furthermore, assume the put costs $60. Then the combination of

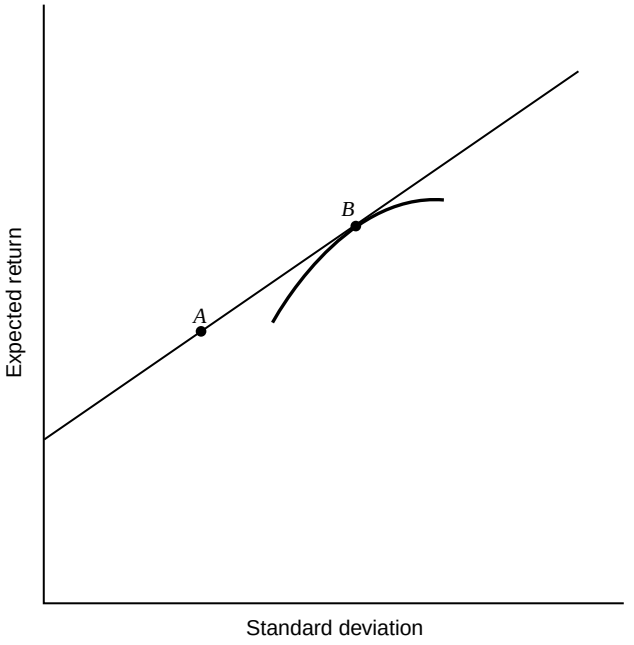


Figure 23.7 The efficient frontier.

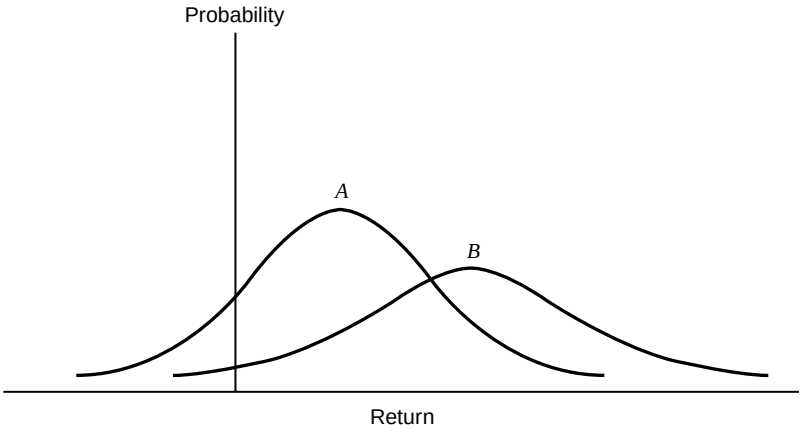


Figure 23.8 Distribution of returns with various amounts in the risky portfolio.

the portfolio and put results in a return distribution such as the one shown by the dashed curve in Figure 23.9. The new distribution of returns is a result of two influences. If the Standard and Poor's (S&P) index ends up above 1,550, the put expires worthless and return is lowered by the cost of the put $\left(\frac{60}{1500} \text{ or } 4\%\right)$. If the S&P index ends up below 1,550, the put is exercised and the return is $\left(\frac{1500}{1500} - 1\right)$ minus the cost of the put, or -0.67%. Thus the effect on the return distribution of buying a put is to lower high returns and eliminate low returns. The worst outcome the investor could incur in our example is a return of -0.67%. This is true no matter how badly the stock market did. This modification of the return

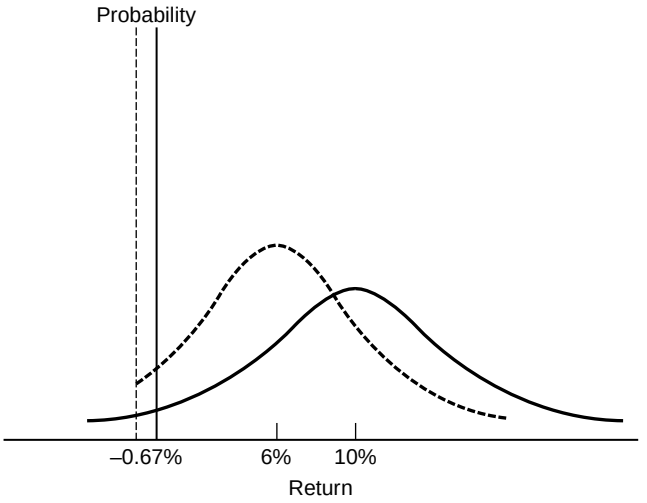


Figure 23.9 The effect of puts on the return distribution.

distribution is not possible with a fixed combination of Treasury bills and a security portfolio.[21]

Another way that options are frequently used to modify the distribution of returns is to sell calls on stocks that are already owned in a portfolio. This gives some immediate income to the investor but has an opportunity cost in that the investor has sold off the right to receive a high potential payoff from the stock. Some investors (e.g., mutual funds), who have a target price above which they intend to close out a position in a stock, find that by selling calls at this price, they can gain extra revenue while following their intended course of action. Some investors who follow this covered call writing strategy are less rational in that they have failed to consider the opportunity of high returns they give up for receiving the income from writing calls. The major use of options is to modify the return distribution in ways unattainable with fixed combinations of other assets.

Betting on Information

Investors receive a great deal of information about the prospective fortunes of company's shares. Information about the company's return can be utilized to buy or potentially short sell a company's stock. One type of information that is not easily utilized with nonoption strategies is information concerning the stock's variability. If an investor believes the company has undergone a large increase in risk, the investor might wish to sell stock if it is already owned, but there is no way with nonoption strategies to utilize this information to justify a purchase or short sale of the stock. Examining the option pricing formulas presented earlier, however, shows that the value of an option is directly related to the stock's underlying volatility. If the investor believes that the volatility of the company will increase dramatically and other investors have yet to discover this increase, then the purchase of options is a way to utilize this information. Thus options are a convenient way to attempt to profit from information about a security's variance.

[21]As discussed earlier, dynamically changing the mix of T-bills and risky assets can create a portfolio with the payoff pattern of a put plus a risky portfolio.

Advanced Uses

There are other ways of employing options that depend on the ability to combine options with other securities to create a portfolio with identical characteristics to yet a third type of security. These are the security equivalencies discussed earlier. For example, combining options plus the underlying security with changing proportions of each can create a portfolio that has the same characteristics as a Treasury bill. If options are mispriced, this allows lending (or borrowing) at more attractive rates than the market. Furthermore, if options are fairly priced, the portfolio of options and the underlying security allow the ability to borrow at the T-bill rate (ignoring transaction costs). Similarly, combinations of options and T-bills dynamically changed through time can create a security with the characteristic of a short position in the underlying security. Because there are limits to the size of the position an investor can short of the basic security, the use of options allows the investor to circumvent exchange restrictions. Finally, it has been argued that because transaction costs are so low in options markets, and because options in conjunction with other securities can create new securities, options may be a less expensive way to buy the created security. All of the uses involve changing the mix of options and a second security over time to create a portfolio with a return pattern like a third security. This involves transaction costs. The value of options for these purposes when transaction costs are included needs to be examined.

CONCLUSION

In this chapter we have examined the characteristics and valuation of contingent claim contracts. The development of a set of models for pricing contingent claims is a fairly recent and important contribution. We have explored the theory behind these models and their use in valuing options. In addition, we have shown how such models can be used to develop estimates of the variance of the return on the stocks against which they represent a claim. This may be an important input to portfolio management models.

APPENDIX A

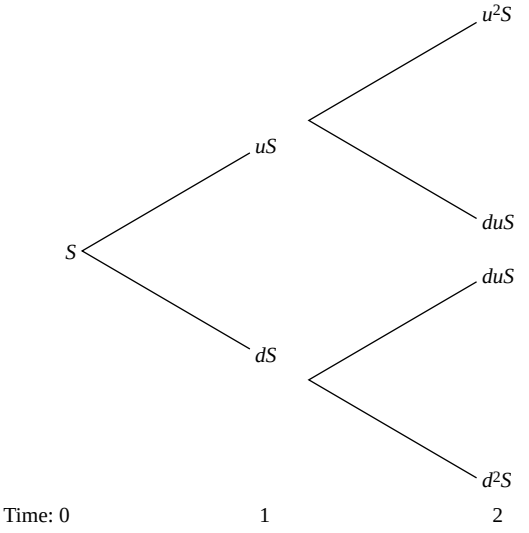
DERIVATION OF THE BINOMIAL FORMULA

In the text we showed that with one period to go, the value of the call was

$$C = \frac{PC_u + (1-P)C_d}{r} \tag{A.1}$$

Now consider the possibilities with two periods to go. These are represented in the following diagram, where

1. C_{u^2} is the value at expiration if there are two up movements in the stock = maximum $[u^2S - E, 0]$
2. C_{ud} is the value at expiration if there is one up and one down movement in the share price = maximum $[udS - E, 0]$
3. C_{d^2} is the value at expiration if there are two down movements in the share price = maximum $[d^2S - E, 0]$



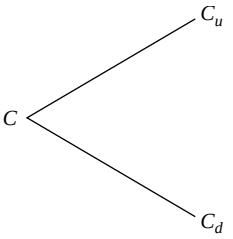
Applying Equation (A.1) we can determine the value of the calls at period 1 if the share price is uS at period 1 as

$$C_u = \frac{PC_{u^2} + (1-P)C_{ud}}{r}$$

Again, by applying Equation (A.1) we can determine the value at period 1 if the share price is dS at period 1 as

$$C_d = \frac{PC_{ud} + (1-P)C_{d^2}}{r}$$

Now consider period 0. Knowing the value at period 1, we can act as if there is only one period to go. This is shown as follows:



(*a*)

Applying Equation (A.1) again yields

$$C = \frac{P\dfrac{PC_{u^2} + (1-P)C_{ud}}{r} + (1-P)\dfrac{PC_{ud} + (1-P)C_{d^2}}{r}}{r}$$

Simplifying,

$$C = \frac{P^2C_{u^2} + 2P(1-P)C_{ud} + (1-P)^2C_{d^2}}{r^2} \qquad \text{(A.2)}$$

In exactly the same way we can derive the formula for the three-period case. The possible movements of the share price are shown in Figure 23.10.

Notice in this case that two periods before the expiration date the stock price is either *uS* or *dS* instead of *S*, as it was in the two-period example. If it is *uS*, then from Equation (A.2) the value at time 1 is simply

$$C_u = \frac{P^2 C_{u^3} + 2P(1-P)C_{u^2d} + (1-P)^2 C_{ud^2}}{r^2}$$

If the price of the stock were *dS* in period 1, then

$$C_d = \frac{P^2 C_{u^2d} + 2P(1-P)C_{ud^2} + (1-P)^2 C_{d^3}}{r^2}$$

where $C_{u^n d^R}$ = the value expiration if there are *n* up movements and *R* down movements. Applying Equation (A.1) yields the value of the call at time zero; we have

$$C = \frac{P\dfrac{P^2 C_{u^3} + 2P(1-P)C_{u^2d} + (1-P)^2 C_{ud^2}}{r^2} + (1-P)\dfrac{P^2 C_{u^2d} + 2P(1-P)C_{ud^2} + (1-P)^2 C_{d^3}}{r^2}}{r}$$

Simplifying,

$$C = \frac{P^3 C_{u^3} + 3P^2(1-P)C_{u^2d} + 3P(1-P)^2 C_{ud^2} + (1-P)^3 C_{d^3}}{r^3}$$

Recalling the determination of the value of the call at the horizon and examining the form of the preceding equations shows that for *n* periods before the horizon, the value of the call is

$$C = \frac{\left[\displaystyle\sum_{j=0}^{n} \frac{n!}{j!(n-j)!} P^j (1-P)^{n-j} \max\left[0, u^j d^{n-j} S_0 - E\right]\right]}{r^n}$$

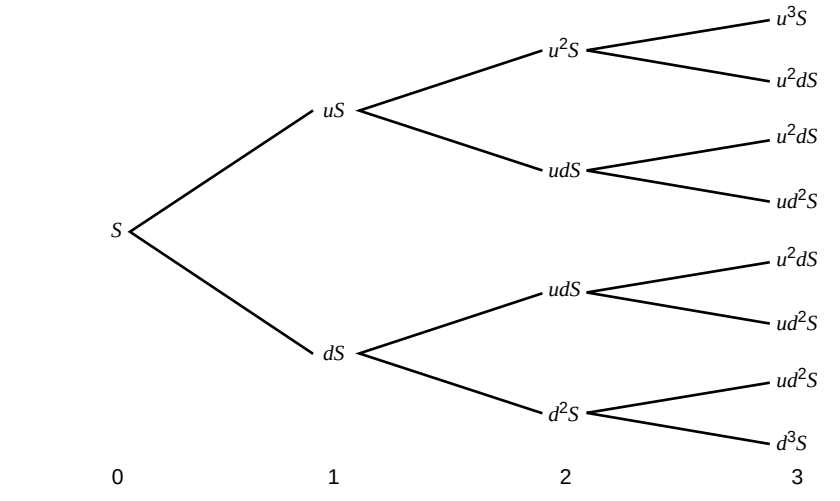


Figure 23.10 Stock price paths.

We can simplify the expression by defining a as the minimum number of up movements necessary for it to pay to exercise the option at the expiration date. For sequences with fewer than a up movements, the call will not be exercised and the value at expiration will be zero. Thus the summation need start only at a. Furthermore, for more than a up movements we know that exercise pays. Thus, when the lower limit on the summation is a, the maximum can be rewritten as $u^j d^{n-j} S_0 - E$.

With these changes we have

$$C = \frac{\sum_{j=a}^{n} \frac{n!}{j!(n-j)!} P^j (1-P)^{n-j} \left(u^j d^{n-j} S_0 - E\right)}{r^n}$$

Rearranging,

$$C = S_0 \left[\sum_{j=a}^{n} \frac{n!}{j!(n-j)!} \frac{(Pu)^j \left[(1-P)d\right]^{n-j}}{r^n} \right] - Er^{-n} \left[\sum_{j=a}^{n} \frac{n!}{j!(n-j)!} P^j (1-P)^{n-j} \right]$$

The second expression in brackets is the binomial formula with R serving the role of a probability and can be represented as $B[a,n,P]$. The first expression in brackets also turns out to be a binomial formula. To see this, first write part of it as

$$\frac{(Pu)^j \left[(1-P)d\right]^{n-j}}{r^n} = \left(\frac{Pu}{r}\right)^j \left(\frac{(1-P)d}{r}\right)^{n-j}$$

Define P' as $(Pu)/r$. Then, if $1 - P' = (1 - P)\, d/r$, we would have a binomial formula with P' serving the role of probability. A little algebra demonstrates that this is appropriate. Recall $P = (r - d)/(u - d)$. Thus

$$\begin{aligned} 1 - P' &= 1 - \frac{Pu}{r} = 1 - \frac{u(r-d)}{r(u-d)} = 1 - \frac{ur - ud}{ur - dr} = \frac{ur - dr - ur + ud}{ur - dr} \\ &= \frac{d}{r}\left[\frac{u-r}{u-d}\right] = \frac{d}{r}\left[\frac{u-d-r+d}{u-d}\right] = \frac{d}{r}\left[1 - \frac{r-d}{u-d}\right] \\ &= \frac{d}{r}(1-P) \end{aligned}$$

This last expression is what we wanted to show. Thus the first term in the brackets has the form

$$\sum_{j=a}^{n} \frac{n!}{j!(n-j)!} P'^j (1-P')^{n-j}$$

This can be represented as $B[a,n,P']$. Substituting the two expressions for binomials in the basic equation for a call yields

$$C = S_0 B[a,n,P'] - Er^{-n} B[a,n,P]$$

APPENDIX B

DERIVATION OF THE BLACK–SCHOLES FORMULA

The derivation of the Black–Scholes formula starts out in a similar manner to the derivation of the binomial formula. First, a portfolio is constructed that has the same return, no

matter how well the stock performs. This portfolio, as in the case of the binomial formula, consists of writing a call and buying the stock. For simplicity, consider buying one share of stock. Then it can be shown that the amount of calls to write is 1 divided by the change in the value of the call, with a unit change in the value of the stock.

The following example will clarify this. Assume that the call changes by one-half of the amount of the stock change. Thus the rule just described says to write two calls. If the stock increased by \$1, the ownership of the stock would cause an increase of \$1 in the value of the hedge. However, if two calls are written, each call should increase in value by \$0.50 or the two calls by \$1. Because the hedge involves writing of two calls, this causes a loss of \$1 in the value of the hedge.

Thus the portfolio value is unchanged by a change in the share price. Such a riskless portfolio should yield the riskless rate of interest. Let

V_H be the initial value (cost) of the hedge

S be the market price of a share of stock

C be the value of a call

Q_s be the quantity of stock owned

Q_c be the quantity of calls owned

r be the riskless rate of interest

Then the value of the hedge is

$$V_H = Q_S S + Q_C C \tag{B.1}$$

and the change in the value of the hedge is

$$dV_H = Q_S dS + Q_C dC$$

This hedge is riskless and thus should yield the riskless rate of interest per each unit of time. Thus

$$rV_H dt = Q_S dS + Q_C dC$$

Substituting for V_H from Equation (B.1), and recalling that if the hedge is formed in terms of writing calls,

$$Q_S = +1 \qquad \text{and} \qquad Q_C = \frac{-1}{\partial C/\partial S}$$

we have

$$r\left[S - \frac{C}{\partial C/\partial S}\right]dt = dS - \frac{1}{\partial C/\partial S}dC$$

Rearranging,

$$dC = \frac{\partial C}{\partial S}dS - r\frac{\partial C}{\partial S}\left[S - \frac{C}{\partial C/\partial S}\right]dt = \frac{\partial C}{\partial S}dS - rS\frac{\partial C}{\partial S}dt + rCdt \tag{B.2}$$

What is required next is a model of stock price and call price changes. The assumption that Black and Scholes make is that the instantaneous change in stock price follows a normal distribution:

$$\frac{dS}{S} = \mu dt + \sigma dZ$$

Variable μ is the instantaneous expected return, σ is the instantaneous standard deviation, and *dZ* is the zero mean unit standard deviation normally distributed variate. Given the stock price process described before, the change in the call price is well known from theorems in stochastic calculus:[22]

$$dC = \frac{\partial C}{\partial S} dS + \frac{\partial C}{\partial t} dt + \frac{1}{2} \frac{\partial^2 C}{\partial S^2} \sigma^2 S^2 dt$$

This expression should look somewhat familiar. The first two terms on the right-hand side are the terms that would be obtained in standard calculus if you take a total derivative of the value of a call. The last term arises because of the stochastic element in *S*. Substituting for *dC* in Equation (B.2) yields

$$\frac{\partial C}{\partial S} dS + \frac{\partial C}{\partial t} dt + \frac{1}{2} \frac{\partial^2 C}{\partial S^2} \sigma^2 S^2 dt = \frac{\partial C}{\partial S} dS - rS \frac{\partial C}{\partial S} dt + rC dt$$

Subtracting the term $(\partial C/\partial S)\, dS$ from each side, noting that there is a *dt* in each remaining term and thus that it can be eliminated by dividing by *dt*, and rearranging yields

$$\frac{\partial C}{\partial t} = rC - rS \frac{\partial C}{\partial S} - \frac{1}{2} \frac{\partial^2 C}{\partial S^2} \sigma^2 S^2 \tag{B.3}$$

This is a differential equation. At the horizon, the value of the call is

$$C = \begin{matrix} S - E & S > E \\ 0 & S \leq E \end{matrix}$$

Solving the differential equation and using the value at the horizon as the boundary condition yields the expression shown in the text.

QUESTIONS AND PROBLEMS

1. A registered representative recently advised one of his clients to sell calls on all the stock he owned. He explained that the client wouldn't lose money but would benefit by what he got paid for the call. Sounds foolproof. What's wrong?

2. Consider the purchase of a combination of two puts and a call. Assume that the call costs $5, the put costs $6, and the exercise price for the put or call is $50. Plot the profit versus the stock price at the expiration date.

3. Consider two calls, one with an exercise price of $40 and one with an exercise price of $45. Assume that the call with the $40 exercise price sells for $8 and the call with the $45 exercise price sells for $5. Assume that they have the same expiration date. Consider the strategy of issuing two $45 calls and purchasing one $40 call. Plot the profit versus the share price at the expiration date.

4. Assume the binomial pricing model. Assume that the share price is $50, the exercise price is $60, $u = 1.2$, $d = 0.9$, $r = 1.1$, and $N = 10$. What is the value of α? What is the call value?

[22]The equation follows from Ito's Lemma. See Black and Scholes (1973).

5. Determine the value of the following call using the Black–Scholes model. The stock currently sells for $95, and the instantaneous standard deviation of the stock's return is 0.6. The call has an exercise price of $105 and has eight months to go before expiration. The continuously compounded riskless rate of interest is 8%.

BIBLIOGRAPHY

1. Arditti, Fred D., and John, Kose. "Spanning the State Space with Options," *Journal of Financial and Quantitative Analysis*, **XV,** No. 1 (March 1980), pp. 1–10.
2. Baesel, Jerome B., Shows, George, and Thorp, Edward. "The Cost of Liquidity Services in Listed Options: A Note," *Journal of Finance*, **38,** No. 3 (June 1983), pp. 989–996.
3. Bailey, Warren. "An Empirical Investigation of the Market for Comex Gold Futures Options," *Journal of Finance*, **42,** No. 5 (Dec. 1987), pp. 1187–1194.
4. Ball, Clifford A. "Estimation Bias Induced by Discrete Security Prices," *Journal of Finance*, **43,** No. 4 (Sept. 1988), pp. 841–865.
5. Ball, Clifford A., and Torous, Walter N. "A Simplified Jump Process for Common Stock Returns," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1983), pp. 53–66.
6. ——. "Bond Price Dynamics and Options," *Journal of Financial and Quantitative Analysis*, **XVIII,** No. 4 (Dec. 1983), pp. 517–532.
7. ——. "On Jumps in Common Stock Prices and Their Impact on Call Pricing," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 155–174.
8. ——. "Futures Options and the Volatility of Futures Prices," *Journal of Finance*, **41,** No. 4 (Sept. 1986), pp. 857–870.
9. Ball, Clifford A., Torous, Walter N., and Tschoegel, Adrian E. "An Empirical Investigation of the EOE Gold Options Market," *Journal of Business Finance*, **9,** No. 1 (March 1985), pp. 101–113.
10. Barone-Adesi, Giovanni, and Whaley, Robert E. "Efficient Analytic Approximation of American Option Values," *Journal of Finance*, **42,** No. 2 (June 1987), pp. 301–320.
11. Beckers, Stan. "On the Efficiency of the Gold Options Market," *Journal of Business Finance*, **8,** No. 3 (Sept. 1984), pp. 459–470.
12. Benninga, Simon, and Blume, Marshall. "On the Optimality of Portfolio Insurance," *Journal of Finance*, **40,** No. 5 (Dec. 1985), pp. 1341–1352.
13. Bhattacharya, Mihir. "Empirical Properties of the Black–Scholes Formula under Ideal Conditions," *Journal of Financial and Quantitative Analysis*, **XV,** No. 5 (Dec. 1980), pp. 1081–1106.
14. Bhattacharya, Sudipto. "Notes of Multiperiod Valuation and the Pricing of Options," *Journal of Finance*, **36,** No. 1 (March 1981), pp. 163–181.
15. Bick, Avi. "Producing Derivative Assets with Forward Contracts," *Journal of Financial and Quantitative Analysis*, **23,** No. 2 (June 1988), pp. 153–160.
16. Black, Fischer, and Scholes, Myron. "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, **81,** No. 3 (May/June 1973), pp. 637–654.
17. Blomeyer, Edward C., and Johnson, Herb. "An Empirical Examination of the Pricing of American Put Options," *Journal of Financial and Quantitative Analysis*, **23,** No. 1 (March 1988), pp. 13–22.
18. Bookstaber, Richard, and Clarke, Roger. "Problems in Evaluating the Performance of Portfolios with Options," *Financial Analyst Journal*, **41,** No. 1 (Jan./Feb. 1985), pp. 48–62.
19. Boyle, Phelim. "Options: A Monte Carlo Approach," *Journal of Financial Economics*, **4,** No. 3 (May 1977), pp. 323–338.
20. Boyle, Phelim, and Ananthanarayanan, A. L. "The Impact of Variance Estimation in Option Valuation Models," *Journal of Financial Economics*, **5,** No. 3 (Dec. 1977), pp. 375–387.
21. Boyle, Phelim P., and Emanuel, David. "Discretely Adjusted Option Hedges," *Journal of Financial Economics*, **8,** No. 3 (Sept. 1980), pp. 259–282.
22. Bracken, Jerome. "Models for Call Option Decisions," *Financial Analysts Journal*, **24,** No. 5 (Sept.–Oct. 1968), pp. 149–151.

23. Breeden, Douglas, and Litzenberger, Robert. "Prices of State-Contingent Claims Implicit in Option Price," *Journal of Business*, **51,** No. 4 (Oct. 1978), pp. 621–651.
24. Brennan, Michael J. "A Theory of Price Limits in Futures Markets," *Journal of Financial Economics*, **16** (1986), pp. 213–233.
25. Brennan, Michael, and Schwartz, Edwardo. "The Valuation of American Put Options," *Journal of Finance*, **XXXII,** No. 2 (May 1976), pp. 449–462.
26. Brennan, Michael J., and Solanki, R. "Optimal Portfolio Insurance," *Journal of Financial and Quantitative Analysis*, **XVI,** No. 3 (Sept. 1981), pp. 279–300.
27. Brennan, Michael J., Schwartz, Eduardo S., Grossman, Sanford J., and Vila, Jean-Luc. "Portfolio Insurance and Financial Market Equilibrium; Portfolio Insurance in Complete Markets: A Note," *Journal of Business*, **62,** No. 4 (Oct. 1989), pp. 455–472.
28. Brenner, Menachem, and Galai, Dan. "On Measuring the Risk of Common Stocks Implied by Options Prices: A Note," *Journal of Financial and Quantitative Analysis*, **19,** No. 4 (Dec. 1984), pp. 403–412.
29. Brenner, Menachem, and Galai, Dan. "Implied Interest Rates," *Journal of Business*, **59,** No. 3 (July 1986), pp. 493–507.
30. Brenner, Menachem, Courtadon, Georges, and Subrahmanyam, Marti. "Options on the Spot and Options on Futures," *Journal of Finance*, **40,** No. 5 (Dec. 1985), pp. 1303–1317.
31. Butler, J. S., and Schacter, Barry. "Unbiased Estimation of the Black/Scholes Formula," *Journal of Financial Economics*, **15** (1985), pp. 341–357.
32. Conover, James A., and Dubofsky, David A. "Efficient Selection of Insured Currency Positions: Protective Puts vs. Fiduciary Calls," *Journal of Financial and Quantitative Analysis*, **30,** No. 2 (June 1995), pp. 295–312.
33. Courtadon, George. "The Pricing of Options on Default-Free Bonds," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 1 (March 1982), pp. 75–100.
34. ——. "A More Accurate Finite Difference Approximation for the Valuation of Options," *Journal of Financial and Quantitative Analysis*, **XVIII,** No. 5 (Dec. 1982), pp. 697–700.
35. Cho, D. Chinyung, and Frees, Edward W. "Estimating the Volatility of Discrete Stock Prices," *Journal of Finance*, **43,** No. 2 (June 1988), pp. 451–466.
36. Cox, Stephen, and Ross, Stephen. "A Survey of Some New Results in Financial Option Pricing Theory," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 383–402.
37. ——. "The Valuation of Options for Alternative Stochastic Processes," *Journal of Financial Economics*, **3,** No. 112 (Jan.–March 1976), pp. 145–166.
38. Dietrich-Campbell, Bruce, and Schwartz, Eduardo. "Valuing Debt Options: Empirical Evidence," *Journal of Financial Economics*, **16** (1986), pp. 321–343.
39. Dimson, Elroy. "Instant Option Valuation," *Financial Analysts Journal*, **33,** No. 3 (May–June 1977), pp. 62–69.
40. ——. "Option Valuation Nomograms," *Financial Analysts Journal*, **33,** No. 6 (Nov.–Dec. 1977), pp. 71–74.
41. Emanuel, David. "A Theoretical Model for Valuing Preferred Stock," *Journal of Finance*, **38,** No. 4 (Sept. 1983), pp. 1133–1155.
42. Eunine, Jeremy, and Rudd, Andrew. "Index Options: The Early Evidence," *Journal of Finance*, **40,** No. 3 (July 1985), pp. 743–756.
43. Eyton, T. Hanam, and Harpaz, Giora. "The Pricing of Futures and Options Contracts on the Value Line Index," *Journal of Finance*, **41,** No. 4 (Sept. 1986), pp. 843–855.
44. Figlewski, Stephen. "Forecasting Volatility," *Financial Markets and Instruments*, **6,** No. 1 (1997), pp. 1–88.
45. Finucane, Thomas J. "Black–Scholes Approximations of Call Option Prices with Stochastic Volatilities: A Note," *Journal of Financial and Quantitative Analysis*, **24,** No. 4 (Dec. 1989), pp. 527–532.
46. Fischer, Stanley. "Call Option Pricing When the Exercise Price Is Uncertain, and the Valuation of Index Bonds," *Journal of Finance*, **XXXIII,** No. 1 (March 1978), pp. 169–176.

47. French, Dan W. "The Weekend Effect on the Distribution of Stock Prices: Implications for Option Pricing," *Journal of Financial Economics*, **13,** No. 4 (Dec. 1984), pp. 547–560.
48. Galai, Dan. "Tests of Market Efficiency of the Chicago Board Options Exchange," *Journal of Business*, **50,** No. 2 (April 1977), pp. 167–197.
49. ——. "On the Boness and Black–Scholes Models for Valuation of Call Options," *Journal of Financial and Quantitative Analysis*, **XII,** No. 1 (March 1978), pp. 15–27.
50. Galai, Dan, and Masulis, R. "The Option Pricing Model and the Risk Factor of Stock," *Journal of Financial Economics*, **13,** No. 1/2 (Jan.–March 1976), pp. 53–81.
51. Garman, Mark B. "The Duration of Option Portfolios," *Journal of Financial Economics*, **14** (1985), pp. 309–315.
52. Geske, Robert. "The Pricing of Options with Stochastic Dividend Yield," *Journal of Finance*, **XXXIII,** No. 2 (May 1978), pp. 617–625.
53. Geske, Robert, and Roll, Richard. "On Valuating American Call Options with the Black–Scholes European Formula," *Journal of Finance*, **39,** No. 2 (June 1984), pp. 443–456.
54. Geske, Robert, and Johnson, H. E. "The American Put Option Valued Analytically," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1511–1524.
55. Geske, Robert, and Shastri, Kuldeep. "Valuation by Approximation: A Comparison of Alternative Option Valuation Techniques," *Journal of Financial and Quantitative Analysis*, **XX,** No. 1 (March 1985), pp. 45–72.
56. Geske, Robert, and Shastri, Kuldeep. "The Early Exercise of American Puts," *Journal of Business Finance*, **9,** No. 2 (June 1985), pp. 207–219.
57. Gould, J. P., and Galai, Dan. "Transactions Costs and the Relationship between Put and Call Prices," *Journal of Financial Economics*, **1,** No. 2 (July 1974), pp. 105–130.
58. Gultekin, N. Bulent, and Rogalski, Richard J. "Government Bond Returns, Measurement of Interest Rate Risk, and the Arbitrage Pricing Theory," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 43–62.
59. Halpern, Paul J., and Turnbull, Stuart M. "Empirical Tests of Boundary Conditions for Toronto Stock Exchange Option," *Journal of Finance*, **40,** No. 2 (June 1985), pp. 481–500.
60. Harrison, Michael J., Pitbladdo, Richard, and Schaefer, Stephen M. "Continuous Price Process in Frictionless Markets Have Infinite Variation," *Journal of Business*, **57,** No. 3 (Oct. 1984), pp. 353–365.
61. Hausman, W. H., and White, W. L. "Theory of Option Strategy under Risk Aversion," *Journal of Financial and Quantitative Analysis*, **111,** No. 3 (Sept. 1968), pp. 343–358.
62. Heath, David C., and Jarrow, Robert A. "Arbitrage, Continuous Trading and Margin Requirements," *Journal of Finance*, **41,** No. 5 (Dec. 1987), pp. 1129–1142.
63. Hilliard, Jimmy, and Leitch, Robert. "Analysis of the Warrant Hedge in a Stable Paretion Market," *Journal of Financial and Quantitative Analysis*, **XII,** No. 1 (March 1977), pp. 85–103.
64. Ho, Thomas S. Y., and Macris, Richard G. "Dealers Bid-Ask Quotes and Transaction Prices: An Empirical Study of Some AMEX Options," *Journal of Finance*, **39,** No. 1 (March 1984), pp. 23–46.
65. Hull, John. *Options Futures and Other Derivative Securities*. (Englewood Cliffs, NJ: Prentice Hall, 2001).
66. Hull, John, and White, Alan. "The Pricing of Options on Assets with Stochastic Volatilities," *Journal of Finance*, **42,** No. 2 (June 1987), pp. 281–300.
67. Jagannathan, Ravi. "Call Options and the Risk of Underlying Securities," *Journal of Financial Economics*, **13** (1984), pp. 425–434.
68. Jarrow, Robert, and Rudd, Andrew. "Approximate Option Valuation for Arbitrary Stochastic Processes," *Journal of Financial Economics*, **10,** No. 3 (Nov. 1982), pp. 347–370.
69. Johnson, H. E. "An Analytic Approximation for the American Put Price," *Journal of Financial and Quantitative Analysis*, **XVIII,** No. 1 (March 1983), pp. 141–162.
70. Johnson, Herb, and Stulz, Rene. "The Pricing of Options with Default Risk," *Journal of Finance*, **42,** No. 2 (June 1987), pp. 267–280.
71. Jones, E. Philip. "Option Arbitrage and Strategy with Large Price Changes," *Journal of Financial Economics*, **10,** No. 4 (March 1984), pp. 91–114.

72. Kassouf, Sheen. "Warrant Price Behavior—1945 to 1964," *Financial Analysts Journal*, **24,** No. 1 (Jan.–Feb. 1968), pp. 123–126.
73. Klemkosky, Robert C., and Resnick, Bruce G. "Put-Call Parity and Market Efficiency," *Journal of Finance*, **34,** No. 5 (Dec. 1979), pp. 1141–1157.
74. ——. "An Ex Ante Analysis of Put-Call Parity," *Journal of Financial Economics*, **8,** No. 4 (Dec. 1980), pp. 363–378.
75. Latane, Henry, and Rendleman, Richard. "Standard Deviations of Stock Price Ratios Implied on Option Prices," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 369–381.
76. Leabo, Dick, and Rogalski, Richard. "Warrant Price Movements and the Efficient Market Model," *Journal of Finance*, **XXX,** No. 1 (March 1975), pp. 163–177.
77. Leland, Hayne E. "Who Should Buy Portfolio Insurance?" *Journal of Finance*, **35,** No. 2 (May 1980), pp. 581–594.
78. ——. "Option Pricing and Replication with Transactions Costs," *Journal of Finance*, **40,** No. 5 (Dec. 1985), pp. 1283–1301.
79. Levy, Haim. "Upper and Lower Bound of Put and Call Option Value: Stochastic Dominance Approach," *Journal of Finance*, **40,** No. 4 (Sept. 1985), pp. 1197–1217.
80. Litzenberger, Robert, and Sosin, Howard. "The Theory of Recapitalization and the Evidence of Dual Purpose Funds," *Journal of Finance*, **XXXII,** No. 5 (Dec. 1977), pp. 1433–1455.
81. Lo, Andrew W. "Semi-Parametric Upper Bounds for Option Prices and Expected Payoffs," *Journal of Financial Economics*, **19** (1987), pp. 373–387.
82. MacBeth, James C., and Merville, Larry J. "Tests of the Black–Scholes and Cox Call Option Valuation Models," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 285–300.
83. Manaster, Steven, and Rendleman, Richard J., Jr. "Option Prices as Predictors of Equilibrium Stock Prices," *Journal of Finance*, **37,** No. 4 (Sept. 1982), pp. 1043–1058.
84. Margrabe, William. "The Value of an Option to Exchange One Asset for Another," *Journal of Finance*, **XXXIII,** No. 1 (March 1978), pp. 177–198.
85. McDonald, Robert, and Siegel, Daniel. "Option Pricing When the Underlying Asset Earns a Below-Equilibrium Rate of Return: A Note," *Journal of Finance*, **39,** No. 1 (Mar. 1984), pp. 261–266.
86. McGuigan, James, and King, William. "Security Option Strategy under Risk Aversion: An Analysis," *Journal of Financial and Quantitative Analysis*, **VIII,** No. 1 (Jan. 1973), pp. 7–15.
87. ——. "Evaluating Alternative Stock Option Timing Strategies," *Journal of Financial and Quantitative Analysis*, **IX,** No. 4 (Sept. 1987), pp. 567–578.
88. Merton, Robert. "The Relationship between Put and Call Option Prices: Comment," *Journal of Finance*, **XXVIII,** No. 1 (March 1973), pp. 183–184.
89. ——. "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science* (Spring 1973), pp. 141–183.
90. ——. "Option Pricing When Underlying Stock Returns Are Discontinuous," *Journal of Financial Economics*, **3,** No. 1/2 (Jan./March 1976), pp. 125–144.
91. ——. "The Impact on Option Pricing of Specification Error in the Underlying Stock Price Returns," *Journal of Finance*, **XXXI,** No. 2 (May 1976), pp. 333–350.
92. Merton, Robert, Scholes, M., and Gladstein, M. "The Returns and Risk of Alternative Call Option Portfolio Investment Strategies." *Journal of Business*, **51,** No. 2 (April 1978), pp. 183–242.
93. Parkinson, Michael. "Empirical Warrant-Stock Relationships," *Journal of Business*, **45,** No. 4 (Oct. 1972), pp. 563–569.
94. ——. "Option Pricing: The American Put," *Journal of Business*, **50,** No. 1 (Jan. 1977), pp. 21–36.
95. Perrakis, Stylianos, and Ryan, Peter J. "Option Pricing Bounds in Discrete Time," *Journal of Finance*, **39,** No. 2 (June 1984), pp. 519–526.
96. Peterson, Richard. "Investor Preferences for Future Straddles," *Journal of Financial and Quantitative Analysis*, **XII,** No. 1 (March 1977), pp. 105–120.
97. Phillips, Susan M., and Smith, Clifford W., Jr. "Trading Costs for Listed Options: The Implications for Market Efficiency," *Journal of Financial Economics*, **8,** No. 3 (June 1980), pp. 179–189.

98. Protopapadakis, Aris, and Stoll, Hans R. "Spot and Futures Prices and the Law of One Price," *Journal of Finance*, **38,** No. 5 (Dec. 1983), pp. 1431–1456.
99. Ramaswamy, Krishna, and Sundaresan, Suresh M. "The Valuation of Options on Futures Contracts," *Journal of Finance*, **40,** No. 5 (Dec. 1985), pp. 1319–1340.
100. Reback, Robert. "Risk and Return in CBOE and AMEX Option Trading," *Financial Analysts Journal*, **31,** No. 4 (July–Aug. 1975), pp. 42–52.
101. Rendleman, Richard, and Bartter, Brit. "Two-State Option Pricing," *Journal of Finance*, **34,** No. 5 (Dec. 1979), pp. 1093–1110.
102. ——. "The Pricing of Options on Debt Securities," *Journal of Financial and Quantitative Analysis*, **XV,** No. 1 (March 1980), pp. 11–24.
103. Ritchken, Peter H. "On Option Pricing Bounds," *Journal of Finance*, **40,** No. 4 (Sept. 1985), pp. 1219–1233.
104. Ritchken, Peter H., and Kuo, Shyanjaw. "Option Bounds with Finite Revision Opportunities," *Journal of Finance*, **43,** No. 2 (June 1988), pp. 301–308.
105. Rubinstein, Mark. "Displaced Diffusion Option Pricing," *Journal of Finance*, **38,** No. 1 (March 1983), pp. 213–217.
106. ——. "A Simple Formula for the Expected Rate of Return of an Option over a Finite Holding Period," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1503–1510.
107. ——. "Nonparametric Tests of Alternative Option Pricing Models Using All Reported Trades and Quotes on the 30 Most Active CBOE Option Classes from August 23, 1976 through August 31, 1978," *Journal of Finance*, **40,** No. 2 (June 1985), pp. 445–480.
108. Rubenstein, Mark, and Cox, John. *Option Markets* (Englewood Cliffs, NJ: Prentice Hall, 1985).
109. Schaefer, Stephen M., and Schwartz, Eduardo S. "Time-Dependent Variance and the Pricing of Bond Options," *Journal of Finance*, **42,** No. 5 (Dec. 1987), pp. 1113–1128.
110. Schmalensee, Richard, and Trippi, Robert. "Common Stock Volatility Expectations Implied by Option Primia," *Journal of Finance*, **XXXIII,** No. 1 (March 1978), pp. 129–148.
111. Schwartz, Eduardo S. "The Pricing of Commodity-Linked Bonds," *Journal of Finance*, **37,** No. 2 (May 1982), pp. 525–538.
112. Sears, R. Stephen, and Trennepohl, Gary L. "Measuring Portfolio Risk in Options," *Journal of Financial and Quantitative Analysis*, **XVII,** No. 3 (Sept. 1982), pp. 391–410.
113. Sharpe, William. *Investments* (Englewood Cliffs, NJ: Prentice Hall, 1978).
114. Shastri, Kuldeep, and Tandon, Kishore. "Valuation of Foreign Currency Options: Some Empirical Tests," *Journal of Financial Quantitative Analysis*, **21,** No. 2 (June 1986), pp. 145–160.
115. Smith, Clifford. "Option Pricing: A Review," *Journal of Financial Economics*, **3,** No. 1/2 (Jan.–March 1976), pp. 3–51.
116. Smith, Keith. "Option Warrant and Portfolio Management," *Financial Analysts Journal*, **24,** No. 3 (May–June 1968), pp. 135–158.
117. Stapleton, R. C., and Subrahmanyam, M. G. "The Valuation of Multivariate Contingent Claims in Discrete Time Models," *Journal of Finance*, **39,** No. 1 (March 1984), pp. 207–228.
118. ——. "The Valuation of Options When Asset Returns Are Generated by a Binomial Process," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1525–1540.
119. Sterk, William E. "Test of Two Models for Valuing Call Options on Stocks with Dividends," *Journal of Finance*, **37,** No. 5 (Dec. 1982), pp. 1229–1238.
120. Stoll, H. R. "The Relationship between Put and Call Option Prices," *Journal of Finance*, **XXIV,** No. 5 (Dec. 1969), pp. 801–824.
121. ——. "Reply," *Journal of Finance*, **XXVIII,** No. 1 (March 1973), pp. 185–187.
122. Stone, Albert. "Option Models," Ph.D. dissertation, New York University (1969).
123. Stulz, René M. "Options on the Minimum or the Maximum of Two Risky Assets: Analysis and Applications," *Journal of Financial Economics*, **X,** No. 2 (July 1982), pp. 161–186.
124. Vu, Joseph D. "An Empirical Investigation of Calls of Non-convertible Bonds," *Journal of Financial Economics*, **16** (1986), pp. 235–265.
125. Weinstein, Mark I. "Bond Systematic Risk and the Option Pricing Model," *Journal of Finance*, **38,** No. 5 (Dec. 1983), pp. 1415–1430.

126. Whaley, Robert E. "Valuation of American Call Options on Dividend-Paying Stocks: Empirical Tests," *Journal of Financial Economics*, **X,** No. 1 (March 1982), pp. 29–58.
127. ——. "Valuation of American Futures Options: Theory and Empirical Tests," *Journal of Finance*, **41,** No. 1 (March 1986), pp. 127–150.
128. Wiggins, James B. "Option Values under Stochastic Volatility: Theory and Empirical Estimates," *Journal of Financial Economics*, **19** (1987), pp. 351–372.

24

The Valuation and Uses of Financial Futures

Forward contracts are commitments entered into by two parties to exchange a specific amount of money for a particular good or service at a specified future time. Although the price is decided upon at the time of the agreement, no cash changes hands at that time. However, either or both parties to the transaction often have to post some funds to guarantee fulfillment of the contract. Forward contracts are a part of everyday life. When one orders a car not in stock from a dealer, one is buying a forward contract for the delivery of a car. The price and description of the car are specified. In this case the delivery date might not be exact. In addition, a deposit is often required to guarantee that the buyer will take delivery and pay the agreed-upon price.

In this chapter we will be primarily concerned with financial futures, though we will say a few words about other types of futures. Financial futures are similar to, but slightly different from, forward contracts. The name *financial future* is very descriptive. *Financial* means that the good to be delivered is a financial instrument (e.g., a stock or bond). The word *future* as opposed to *forward* reflects the fact that on these contracts, profits and losses are computed and settled on a day-to-day basis rather than at the end of the contract. This is called *marking to the market*, and we will have more to say about it shortly. In addition, contracts for financial futures are traded on organized exchanges that set standard terms for the contracts.

This chapter is divided into four sections. In the first section we describe in more detail the characteristics of financial futures. In the second section we show how these contracts can be valued. In the third section we discuss how financial futures can be used in the investment process. Finally, in the fourth section, we briefly discuss commodity futures and commodity funds.

DESCRIPTION OF FINANCIAL FUTURES

A financial futures contract calls for the delivery of either a specific financial instrument or a member of a set of financial instruments at a specific date or during a specific period of time for an agreed-upon price. Financial futures are traded on organized exchanges and have standardized contract terms. The exact terms differ from financial future to financial future. Table 24.1 lists some of the financial instruments on which financial futures are

Table 24.1 Some Underlying Instruments with Financial Futures

Debt Instruments	Equity Instruments	Currencies
30-day Fed funds	S&P 500	Australian Dollar
3 week T-bill	S&P 500 Growth	British Pound
2-, 3-, 5-, and 10-year Treasury note	S&P 500 Value	Brazilian Real
Treasury bond	S&P Mid-Cap 400	Canadian Dollar
Sovereign yield spreads	S&P Small Cap 600	Chinese Renmenbi
CME Barclays U.S. Aggregate Bond Index	NASDAQ 100	Euro FX
Eurodollar	Nikkei 225 (Dollar)	Japanese Yen
10-year Government of Canada Bond	E-Mini Energy Select Sector	Russian Ruble
5-, 7-, 10-, 30-year interest rate swaps		Swedish Krone
Sovereign yield spreads		

Abbreviations: NASDAQ, National Association of Security Traders; S&P, Standard and Poor's.

being or have been traded. We can categorize these instruments as debt instruments, stock indexes, and foreign currencies.

The terms of a futures contract are always specified in detail. These include

1. the amount and type of asset to be delivered—exactly what asset or set of assets must be delivered and in what quantities
2. the delivery date or maturity date—the date or period of time at which the exchange is to be consummated
3. the exact place and process of delivery

In addition, the exchanges often place certain restrictions on trading. For example, they set

1. margin, the amount of funds that must be put up to ensure that each party will follow through with her side of the transaction
2. limits on the size of price changes that can occur within a trading day and the size of positions that can be taken

These restrictions are imposed to ensure orderly markets.

Let's start by discussing the profits or losses from trading futures. Then we will return to an examination of some of the attributes of financial futures that affect their performance.

Profits and Losses from Futures Contracts

Futures contracts are traded on organized exchanges and have prices determined at any moment in time, just as do stocks and bonds. In the next section of this chapter we examine how these prices are determined in the marketplace. For now, let us consider the profit and loss that accrues to the parties to a futures contract. The purchaser of a futures contract is said to be long a contract. The purchaser agrees to take delivery of a certain financial instrument at a certain time. The seller is said to be short the contract; the seller agrees to deliver the instrument at a certain time. We will first examine the case of the purchaser of a specific contract.

Let us consider one contract of government bonds for delivery in 10 days. Government bond contracts are traded in amounts of $100,000 face value. Assume the settlement price series is as shown in Table 24.2. The price is per $1,000. Thus 66 represents $66,000.

For a moment, let us assume that these prices were on a forward contract, rather than on a futures contract. For a forward contract, gains and losses are settled at the maturity

Table 24.2 Cash Flows on a Forward and Futures Contract

Day	Settlement Price	Cash Flow If Long 1 Forward	Cash Flow If Short 1 Forward	Cash Flow If Long 1 Future	Cash Flow If Short 1 Future
−9	66				
−8	67	0	0	+1,000	−1,000
−7	68	0	0	+1,000	−1,000
−6	65	0	0	−3,000	+3,000
−5	64	0	0	−1,000	+1,000
−4	66	0	0	+2,000	−2,000
−3	64	0	0	−2,000	+2,000
−2	68	0	0	+4,000	−4,000
−1	67	0	0	−1,000	+1,000
0	68	+2,000	−2,000	+1,000	−1,000
Total Cash Flow		+2,000	−2,000	+2,000	−2,000

date. At the maturity, time 0, the forward price must be the same as the price for immediate delivery since the forward and spot contract each require immediate delivery of the same instrument. The original buyer of the forward contract has the obligation to buy the bond at $66. At the maturity of the forward contract, bonds cost $68. The profit to the buyer of the contract is $2 times the $1,000, or $2,000. Correspondingly, the seller of the contract is selling a bond at $66 when the market price is $68. This is a loss of $2 times $1,000, or $2,000.

Note that the profit to the purchaser is equal to the loss of the seller. Forwards and futures are zero-sum games; the profits (or loss) of the purchaser plus the loss (or profit) to the seller equals zero.

The cash flow pattern from the viewpoint of a buyer or seller of a futures contract is different from and more complex than the cash flow pattern from a forward contract. This is because, as mentioned earlier, futures contracts are marked to the market on a daily basis. At the close of each trading day, the gain or loss from the price change that occurred over that day is immediately credited or debited to the accounts of the individuals who are long or short. Furthermore, all contracts are rewritten so that the price at which parties are obliged to buy and sell the financial instrument is the price of the future at the close of the day. This repricing is referred to as "marking to the market," and the price used to mark to the market is called the settlement price. The aggregate profit or loss from the contract over the life of the contract would be the same whether it is a forward or a futures contract, namely, $2,000, but the timing of the cash flows is very different.

As just discussed, the person who shorted (wrote) a forward contract with the price behavior displayed in Table 24.2 would have one cash flow of −$2,000 on day 0. The person who wrote a futures contract would have a much more complex pattern. On day −8, the futures price is $67. So the writer of the futures contract would then be debited by $1,000 on day −8. The writer of the futures contract would then be considered to have a contract at $67. The contract is marked to the market. On day −7, futures go to $68. Since the contract is implicitly at $67, the loss is again $1,000. The writer will have the account debited by $1,000, and the price of the contract will be specified at $68. Thus the writer of the futures contract has the series of intermediate cash flows shown in Table 24.2. If the reinvestment rate were zero, an investor would not care whether a future or forward contract was held. However, the potential of receiving cash or having to come up with cash on a daily basis makes futures contracts different from forward contracts.

Some Important Attributes of Futures Contracts

In this section we discuss three aspects of financial futures that can impact their performances. These are margin, limits, and delivery.

Margin To purchase or sell a future is actually to enter into a promise to take a future course of action with associated cash flows over time. This is not a traditional investment because, at the time a futures position is bought or sold, no cash changes hands between the two parties. However, to ensure that the parties can fulfill their obligation, an initial margin or good faith deposit must be made with the broker. The size and terms of the good faith deposit vary from future to future. They are generally related to the size of the contract and the variability in the daily value of the contract. Relating to daily variability makes sense because the purpose of the good faith deposit is to see that contracts are fulfilled and that contracts are adjusted for profits or losses (marking to the market) on a daily basis. Margins for futures are small relative to other types of markets. For example, the initial margin needed to buy a future on $1 million face value of Treasury bills (T-bills) is often about $1,000. Furthermore, the margin can be put up in the form of earning assets such as T-bills or letters of credit. Nowadays, every futures market has a maintenance margin level, usually 75%–80% of the initial margin level. If margin drops below the maintenance level due to marking of the market, then the investor must come up with additional funds to bring the account back to the required margin level. The cash flows needed to do so are called variation margin, and this added margin must be put up in the form of cash. If the investor does not deposit the added margin, the broker can liquidate the position at the going market price. The investor is liable for any shortfall that occurs when his portfolio is liquidated.

Limits Another aspect of futures markets that should be discussed is the existence of limits. Most financial futures markets have limits on the size of the position any investor can take. Of more interest is that they have limits on the size of the price change that is allowed to take place during any day. For example, price moves on the $1 million 90-day T-bill contract are limited to $1,500 per day. When the price moves up or down by that amount during a day, trading essentially stops. What this can mean (and in fact has meant in the futures market for silver, among others) is that a position cannot be closed out during a period of time at any price. There have been periods of time where price has moved down by the limit for a number of days in a row and no one has offered to buy at that price. Thus for a number of days it was literally impossible to sell on organized exchanges. While these limits were imposed to ensure orderly markets, they constitute an added risk to investing in futures markets.

Delivery The delivery options of financial futures contracts are well specified. However, the person who has shorted the future often has several options as to which of several financial instruments to deliver and sometimes an option on exactly what day to deliver. For example, in dealing with the futures on Treasury bonds, the investor who has shorted such a future can deliver any government bond with more than a 15-year maturity and less than a 25-year maturity. There are many bonds in the market at any time that fit the description and hence can be delivered. The amount of any bond that must be delivered to satisfy the contract is well specified. A set of "conversion factors" has been determined to ensure that a delivered bond would have the same yield to maturity or, if callable, yield to first call as a 6% coupon bond. The attempt was to make a large number of bonds equivalent for delivery. However, over most periods of time, these bonds are not

equivalent. Generally, at any point in time, there is a bond that is "cheapest" to deliver. This option to deliver any bond, along with the option to deliver at any point over a short period of time, adds value to the position of the future seller and correspondingly subtracts it from the value of the buyer. However, the seller always wants to deliver the cheapest bond. Historically, this bond has been stable over long periods of time. Even in turbulant times the prices of the cheapest to delivery and the former cheapest to deliver are close. Thus the reader should not overemphasize the value of this option.

Not all financial futures are settled by delivery of an asset. Some (e.g., stock index futures) are always settled for cash. In this case, the final settlement price of the futures contract is set equal to the market price of the underlying assets on the last trading day of the contract.

Delivery of an asset rarely takes place, even for those contracts that are theoretically settled by delivery of an asset. Almost all futures positions are settled by an offsetting trade rather than by delivery. For example, a buyer of a June Treasury bond contract can close out that position at any time by selling a June Treasury bond contract. Less than 1% of all futures contracts traded are settled by delivery of the underlying asset.

VALUATION OF FINANCIAL FUTURES

The valuation of financial futures is greatly simplified by understanding the relationship between futures prices and the current (or spot) price of the underlying financial instrument. As we will show, a particular relationship must exist, for if it fails to hold, then an immediate riskless profit could be made. Because there are many individuals continuously looking for opportunities to profit from just such a failure, these basic relationships are reasonably descriptive of real markets. We will examine the relationship between spot and future prices for each of the major financial futures. All of the pricing relationships are derived from the ability to hold a financial instrument directly or to create a second instrument with almost identical cash flows by buying or selling futures contracts.

Treasury Bill Futures

Consider an investor who wants to hold a 151-day T-bill. The investor could do so in either of two ways. First, the investor could purchase it directly by simply buying the 151-day T-bill. Alternatively, the investor could purchase it indirectly. The investor could buy a forward contract on a 91-day T-bill for delivery in 60 days. Simultaneously, the investor could purchase a 60-day T-bill that matures for an amount exactly sufficient to take delivery of the forward contract. As we show, the resulting cash flows are identical. Both investments involve an immediate cash outlay and an inflow of the same size in 151 days. Because the future cash flows are the same, the price (initial cash flow) must be the same. Let us look at the two ways to purchase a 151-day T-bill in more detail.

1. *Directly.* Buy a 151-day Treasury bill at a cost of P.
2. *Indirectly.* Buy a forward contract that will lead to delivery in 60 days of a 91-day T-bill. Let us define F as the price the holder of a forward contract must pay to take delivery of bills in 60 days. Buy a 60-day T-bill that will have a value equal to F at the delivery date. This will cost $F/(1 + R)$, where R is the interest rate on a 60-day T-bill.

The cash flows for these two strategies are shown in Table 24.3. Both of these strategies produce identical future cash flows (equivalent to that of holding a 151-day instrument).

Table 24.3 Cash Flows on T-bill and Homemade T-bill Contract

Action	0	60	151
Direct			
Buy 151-day T-bill	P_0		1,000,000
Indirect			
Buy forward contract and take delivery		$-F$	1,000,000
Buy 60-day T-bill	$\underline{F/(1+R)}$	$\underline{+F}$	
Sum	$F/(1+R)$	0	1,000,000

Because they have identical cash flows, they should have the same cost. Thus

$$F/(1+R) = P \tag{24.1}$$

That two identical instruments should sell at the same price is known as the "law of one price."

If the prior relationship does not hold, then there are profit opportunities. The presence of such profit opportunities should reasonably alert investors to try to exploit them and in the process cause the relationship to hold. There are three types of profit opportunities that should force Equation (24.1) to be an equality.

Buy the Cheapest Instrument The real 151-day T-bill and the homemade 151-day T-bill are identical instruments. Anyone who wished to hold a 151-day instrument should buy the least expensive of the two. This will bid up the price of the cheaper and cause the more expensive to decrease in price. To buy either of the two, the investor would incur transaction costs. If there are a sufficient number of investors with a desire to buy a 151-day instrument, the return of the two instruments could be affected only by the difference in transaction costs between the direct and the indirect purchase of a T-bill, and these should be exceedingly small. Thus Equation (24.1) should be extremely accurate.

Swap Assume that the homemade 151-day T-bill is cheaper than the traded bill. In this case, anyone holding the 151-day T-bill would earn an immediate profit equal to the difference in their prices less transaction costs by selling the 151-day T-bill and purchasing the homemade 151-day T-bill. Such a trade will involve transaction costs on the purchase and on the sale. If the return differential is greater than two transaction costs, an alert investor will undertake the swap. This should force Equation (24.1) to be close to an equality. Because transaction costs in the T-bill and futures markets are very small, the equality should be very closely approximated.

Pure Arbitrage The final force causing the law of one price to hold is pure arbitrage. Arbitrage involves selling short the more expensive instrument and using the proceeds of the sale to purchase the cheaper. Since subsequent cash flows are identical, this involves an immediate profit. The transaction costs are on the purchase and sale. In addition, the short seller usually incurs a $\frac{1}{2}$ of 1% cost on the short position. This last force causing the law of one price to hold is the most powerful in the sense that there are a large number of alert arbitrageurs prepared to take advantage of any discrepancies in the market. At the same time, because transaction costs are higher, the difference between a 151-day T-bill and a homemade 151-day T-bill can be larger without it paying to eliminate the differences.

Which of these three profit opportunities sets prices is still an open question. However, any of them should cause the returns of the 151-day T-bill and the homemade 151-day T-bill to be close and Equation (24.1) to be a reasonable equation for pricing forward Treasury bills.

There are several simplifications in our analysis. We used a forward contract in our discussion. However, futures contracts are the contracts that are available to most investors. The reason we used forward contracts was to avoid the intermediate cash flows associated with marking futures contracts to the market. Futures contracts, of course, involve marking to the market. Thus the homemade 151-day T-bill has intermediate cash flows. However, marking to the market does not have much of an impact for T-bill futures. Elton, Gruber, and Rentzler (1984) found that marking to the market affected cash flows on a million dollar T-bill position by an average of only $4. A difference in cash flows due to marking to the market in the range of plus or minus ($31) occurred 75% of the time. To put these numbers in context, recall that T-bill futures are sold in million-dollar denominations, so that over 60 days at 9% interest the total cash flow is close to $15,000. Thus the effect of marking to the market is trivial, and Treasury bill futures can be sensibly treated as if they were forwards. Elton, Gruber, and Rentzler (1984) did an extensive analysis of the difference in returns of the actual bill and homemade bill. There were differences, and any strategy that involves selling futures generally offered the higher return. Although this is evidence that the law of one price does not hold exactly, and the market is not perfectly efficient, the differences were quite small.

Treasury Bond Futures

Treasury bills are government debt of one year or less to maturity. In addition, they are pure discount instruments with no intermediate cash flows. Treasury bills sell for less than their value at maturity. The increase in value from the time of the sale to the maturity provides the return to the investor. Treasury bonds, in contrast, have original maturities longer than one year and provide a periodic coupon payment as well as potential capital appreciation or loss. A futures market exists for Treasury bonds that have at least 15 years to maturity and are either noncallable or, if callable, are not callable for at least 15 years. There are a large number of different government bonds that meet these criteria, and any of them can be delivered to settle a Treasury bond futures contract. The standard bond to be delivered is a 6% coupon bond. Conversion factors have been computed for bonds with different coupon rates. Bonds with a different coupon are worth some fraction (for higher coupon bonds, a fraction greater than one) of the standard 6% bond. When the conversion factors were created, the hope was that there would be many different issues that would be equivalent and could be delivered. In practice, there is generally a single bond that is cheapest to deliver and will be delivered if actual delivery takes place. This particular issue that is cheapest to deliver is fairly stable over time. Thus the bond that would be potentially delivered is fairly well known at the time the futures contract is written. There are two ways an investor who wished to purchase a government bond could do so:

1. *Purchase directly.* The investor purchases the bond at a current spot price, which can be represented as P.
2. *Purchase with delay.* The investor buys a Treasury bond future with a delivery price of F and simultaneously buys a T-bill with a face value of F that matures at the delivery date on the futures contract. The cost of the T-bill is $F/(1 + R)$, where R is the T-bill rate for the time until the future is delivered.

If the bond does not pay interest before the delivery of the future, these are equivalent positions, and if the law of one price holds, they should have the same cost. Thus

$$F/(1+R) = P \tag{24.2}$$

If the Treasury bond has an interest payment before the delivery date, then the price of direct purchase should be reduced by the present value of this payment. If I is the payment and $PV(I)$ is the present value of the payment, then the law of one price implies

$$F/(1+R) = P - PV(I) \tag{24.3}$$

Once again we are ignoring marking to the market. In addition, the foregoing analysis assumes we know which bond will be delivered. Historically, the bond that will be delivered, the so-called cheapest deliverable instrument, has usually remained stable over long periods of time, so that this assumption holds reasonably well in practice. In addition, all of the earlier discussion on what causes the law of one price to hold still follows. Thus Equation (24.2), though a very good estimate of the futures value, should not be expected to hold exactly.

There are other debt instruments with futures markets available. The reader should be able to modify the preceding analysis to value these alternative instruments.

Stock Index Futures

Futures exist on a number of stock market indexes such as the Standard and Poor's (S&P) 500 Index, the S&P 100 Index, the Value Line Index, and the New York Stock Exchange (NYSE) Composite Index. The NYSE Composite is a value-weighted index of all the stocks on the NYSE. Its return is equivalent to the capital appreciation on a portfolio of all stocks listed on the NYSE where the weights in the portfolio are proportional to the market value of the stock (number of shares times price per share). The S&P 100 and 500, as the names imply, have 100 or 500 firms in the index, and these are generally the largest firms on the NYSE. They are also value-weighted indexes. The Value Line Index has a peculiar construction and, given its unimportance in the futures market, will not be discussed further here. The introduction of stock index futures was delayed by the lack of deliverable instruments. What facilitated their introduction was the acceptance of a cash settlement. For most stock index futures, the future is marked to the spot when the future expires. Cash is then transferred at that point in time; no instrument is ever delivered. While there are lots of ways to arrive at the value of stock futures, the easiest way is to assume that an investor looks at the following alternatives: buy an index fund leveraging the position so that the expected cash flow prior to a particular date is zero or buy T-bills and futures so the same purchase of the index fund is accomplished at maturity.

1. *Direct purchase of an index fund, taking action to eliminate intermediate cash flows.* Assume the index fund can be bought for P dollars and the expected dividend on this amount of the fund is D. Let $PV(D)$ be the present value of the expected value of the future dividend stream. Borrow enough money so that the debt is repaid with the dividend. That is, borrow $PV(D)$ and use these borrowed funds to pay for part of the index fund. Thus the amount of cash that must be put up is $P - PV(D)$.
2. *Indirect purchase.* Buy futures that will represent the same amount of the index fund for an amount of money, F, and simultaneously buy a T-bill that will mature at a value F. The T-bill costs $F/(1 + R)$. Because the amount of the index fund purchased is the same, and because the dividend flows are used to repay the borrowing, the cost

must be the same, or

$$F/(1+R) = P - PV(D)$$

or

$$F = P\left[(1+R) - \frac{PV(D)}{P}(1+R)\right] \tag{24.4}$$

There are arbitrageurs who continually monitor this relationship and take action if it is out of line. Who are the arbitrageurs? One group is index fund managers. Brokerage firms continually monitor the relationship between stock index values and stock index futures prices. When the futures are cheap, they offer to buy a part of an index fund and to sell the fund T-bills and futures. Because the stock trading is not based on a belief that individual shares are mispriced, the brokerage firm believes that the shares can be rapidly resold and thus can offer very low transaction costs; $\frac{1}{8}$ of a dollar is not unusual. Likewise, if futures become overpriced, the reverse trade is made. In terms of our earlier discussion of how the law of one price comes about, this is considered a swap. In addition, there are arbitrageurs in the market who have constructed a small portfolio that is highly correlated with the index. This portfolio is bought or sold short, depending on the value of Equation (24.4), with a corresponding action taken in the futures market.

Before leaving this section, it is appropriate to once again emphasize the factors that might cause the formula not to fit exactly. The formula depends on a forecast of dividends. However, these are dividends on an index, so they are relatively easy to forecast. Nevertheless, there is some small amount of risk in dividend forecasts, and this could introduce some added risks when attempting to duplicate the performance of an index fund with futures. The formula also ignores any effect of marking to the market. In the case of stock index futures, and insofar as a stock index is a proxy for a market portfolio in a capital asset pricing model sense, these flows may be correlated with the market and consequently may introduce systematic risk. We now turn to a discussion of foreign currency futures.

Foreign Currency Futures

Futures exist on all the major currencies. Table 24.1 shows a few of the currencies for which futures contracts exist. Once again, two equivalent instruments can be created that allow the valuation of the futures contract. The two equivalent instruments in this case are riskless domestic debt and riskless foreign debt. Foreign riskless debt is held as follows: convert dollars to a foreign currency, for example, pounds. Invest the money in the foreign riskless debt. Guarantee the rate of conversion back to dollars with financial futures. This is accomplished by writing a futures contract converting pounds to dollars at the maturity of the foreign T-bill for an amount equal to the maturity value of the T-bill. Since the conversion to dollars is at a known rate, the foreign investment is riskless. Futures are quoted in number of dollars per pound. Let S be the initial number of dollars that can be bought with one pound. The initial conversion is to convert dollars to pounds. To convert dollars to pounds, we use one over the rate, or $1/S$. For example, if the rate is \$2 per 1£, one dollar is worth half a pound (1/2). Finally let F be the futures price of one pound and R^B be the foreign (British) riskless rate. Consider an investment in British riskless debt. Then the number of pounds bought per dollar invested is $1/S$. The value of the debt at maturity is $(1 + R^B)/S$. Finally, the value at the maturity in dollars is

$$\frac{\left(1+R^B\right)}{S}F$$

and the return is

$$\left[\frac{\left(1+R^B\right)}{S}F\right]-1$$

If the law of one price holds, all riskless debt should have the same return. If R^D is the rate of return on domestic debt, then

$$R^D=\left[\left(1+R^B\right)F/S\right]-1$$

or

$$F=\left[\left(1+R^D\right)/\left(1+R^B\right)\right]S \tag{24.5}$$

Equation (24.5) is known as interest rate parity or more properly covered interest rate parety. Empirically, interest rate parity seems to hold fairly well. The risk element besides marking to the market is a fear of exchange controls. Governments can and do restrict conversion from one currency to another. In addition, governments can tax the returns to foreign investors. This affects the relative return and can be one element of risk insofar as a change in the tax law can occur during the time of the hedge.

In this section, we have discussed the valuation formula for commonly traded financial futures. The same principles should hold for the financial futures we have not discussed here.

THE USES OF FINANCIAL FUTURES

The growth in financial futures trading has been astronomic in recent years. For example, the dollar volume of shares commanded by futures contracts traded on the S&P index on an average day exceeds the dollar volume of direct trading in these shares. The major markets for financial futures are liquid and involve low transaction costs. Transaction costs are only a fraction of those involved in trading the underlying assets commanded by futures contracts. The combination of liquidity with low transaction costs has meant that there are a large variety of uses for financial futures contracts. We will attempt to review only a few of them here. We find it helpful to divide the uses to which financial futures can be put into three categories: hedging, investment management, and investment products. There is overlap between these categories, but they do serve as a useful characterization.

Hedging

The use of financial futures as a hedging mechanism has received the most attention in the financial literature. Hedging refers to the use of financial futures to reduce a type of risk to which the buyer or seller is subject. For example, the corporation about to sell a long-term issue of bonds (or the underwriter of such an issue) can eliminate most of the risk of interest rate movements by selling a future on a like amount of long-term government bonds. By doing so the corporation in essence locks in the current interest rate. If interest rates go up, the corporation will have to pay a higher interest rate to sell its bonds, but it will find that the value of its short position in futures has gone up by a similar amount. Unfortunately, this may not be an exact dollar-for-dollar movement because of basis risk. Basis risk is the risk that the spot price of the firm's corporate bonds and the futures price of government bonds do not move exactly alike. Corporate bonds and government bonds do move in similar but not identical ways over time. For example, when interest rates go up, the price of

both long-term government bonds and long-term corporate bonds go down, but the amounts by which they go down need not be exactly the same because the spread in rates between the two instruments can change. However, the divergence of these rates over time is very small relative to the effect on the prices of either instrument as the level of rates changes.

As another example of hedging, consider a corporate treasurer who expects to receive a large sum of money in three months to invest in Treasury bills. By buying T-bill futures now, he or she can lock up a known rate on T-bills. In fact, if the treasurer takes delivery, the return will be certain.

Finally, consider an investor due to retire who is worried about the value of that portion of the pension fund that is invested in common stocks. By selling futures on a widely diversified portfolio like the S&P index, the investor can hedge away the risk that the stock market will go up or down between now and the time of retirement.

Changing Investment Policy

Financial futures have transaction costs that are dramatically less than those on stocks and bonds. This implies that they are likely to be the preferred way to change the risk exposure of individual assets or categories of assets. In addition, the use of financial futures allows a direct measure of the value added or subtracted by the policy change. Finally, using financial futures allows a wider choice of assets because of an ability to change risk exposure without having to buy and sell the individual assets in the portfolio. The ideas just presented need elaboration. The elaboration is best done with a few examples.

Changing the Market Exposure of a Stock Portfolio Consider a manager of a mutual fund with a particular exposure to changes in market level. Assume for a moment that the beta on the portfolio is 1.5. Thus a 1% move in the market should be expected to lead to a 1.5% change in the rate of return on the portfolio. Assume further that the manager is pessimistic about the future course of the market and wishes to reduce the exposure. Without financial futures the manager would sell high-beta stocks and purchase lower-beta stocks or T-bills with the proceeds. With financial futures the beta on the portfolio can be reduced in an alternative manner. If the manager sells stock index futures, the combination of the existing portfolio and the stock index futures will have a reduced beta. By selling sufficient stock index futures, the manager can reduce the beta to any level desired. Conversely, if the manager wished to increase the beta on the portfolio, financial futures could be purchased.

There are a number of advantages in using futures to control the risk exposure of the portfolio to market fluctuations. First, transaction costs on futures are dramatically lower than transaction costs of selling stock and purchasing T-bills, or of using stock swaps to change the portfolio beta. Second, if the firm feels that it has the ability to select individual stock issues, then changing the market exposure by selling stock and purchasing T-bills reduces the contribution of selection ability to the performance on the overall portfolio.

For example, if a manager who felt that the return on a particular stock portfolio would be 1% above equilibrium reduced its market exposure by being half in T-bills, the return on the full portfolio should be $\frac{1}{2}$ of 1% above equilibrium. By contrast, if futures were used to control risk exposure, the full 1% would generally be earned on the portfolio. Likewise, if the manager controls market risk exposure on the overall portfolio by constraining the beta to be a particular level, then the performance should be reduced if there is forecasting ability, because less promising stocks must be selected in order to maintain a promised risk level. The final advantage of using futures is that they allow a direct evaluation of timing.

Managers often try to vary their market risk exposure because of a belief in their ability to anticipate market moves. If the manager uses futures to time, then profits and losses on the futures position are a direct measure of the manager's timing ability. The use of futures separates performance due to timing from performance due to selection.

Changing Interest Rate Exposure on Bonds In prior chapters we discussed the concept of bond duration. Bond duration is a measure of the sensitivity of a bond portfolio to changes in interest rates. Most bond managers are timers. They forecast the future course of interest rates. If they feel rates will rise more than they had previously anticipated, they shorten the maturity of their portfolio. If they feel rates will fall more than they had anticipated, then the maturity will be lengthened. Transaction costs across bonds can vary dramatically. The transaction costs of very liquid government bonds are substantially less than those of thinly traded government or corporate bonds. Because of this, timers generally hold very liquid governments as a large part of their portfolio. Thus, timers are giving up the greater expected return of corporates and less liquid governments in order to have lower transaction costs.

Futures can accomplish the same purpose without constraining the investor to holding lower-return securities. The duration on a portfolio can be changed by buying or selling futures. If the manager wishes to shorten the duration, then futures are sold. For example, consider the issuance of a one-year future on a Treasury bond when Treasury bonds are held in the portfolio. The Treasury bond could be delivered against the future in a year. Thus the maturity of the bond has switched from long term to one year with a corresponding change in the duration. If the manager wishes to lengthen the maturity of the portfolio, then futures are bought.

There are a number of advantages in using futures to change interest rate exposure. These are the same reasons discussed earlier in using stock index futures. However, they bear repeating here. First, the transaction costs are substantially less. Second, using futures allows the manager to make the selection decision independent of the duration decision. Thus if the manager feels that certain sectors or bonds are especially attractive, these bonds can be selected even if they are illiquid and can be sold only with large transaction costs. Even if the manager does not profess to have selection ability, bonds that are illiquid but promise higher returns (such as corporates) can be selected. The reader should note that the manager can be exposed to basis risk by using futures on instruments that differ from those held in the portfolio to change duration. However, the impact on returns due to instrument types that are not perfectly matched should be small compared to the impact of interest rate changes on portfolios of different durations. Finally, profits and losses on the futures are a direct measure of the timing ability and the value added by the timing ability, if any exists. Many managers profess to have timing ability, but it is difficult to measure. Using futures gives a direct measure.

Before leaving this section one other issue will be briefly discussed. When interest rates were extremely high, the duration on even long-term bonds was fairly short. The problem was that the liabilities of many institutions (e.g., pension funds) had longer durations than even the longest maturity bond. Thus immunization, the matching of the duration on assets and liabilities, was infeasible using bonds alone. Futures can be used to change duration. A mixture of bonds and futures could be constructed with an arbitrarily long duration. Thus in periods of higher interest rates, employing futures was the only way that immunization was possible for many liabilities.

Changing the Bond–Stock Mix Consider a manager of a balanced fund. One of the decisions that must be made is the relative exposure to the stock market and to the bond

market. If the manager decides that the exposure to the stock market should be increased, it should be obvious from the prior sections that this change can be accomplished with futures. Purchasing stock index futures will increase the beta on the stock portfolio and increase the stock market exposure. Selling bond futures decreases the duration and reduces the exposure to interest rate changes. All of the advantages of futures discussed in prior sections still hold for this use of futures. In particular, the use of futures lowers transaction costs, allows security selection to be independent of the market exposure decision, and gives an unambiguous measure of timing ability.

Creating New Products

Futures have been used to create products that could not exist or were inordinately expensive before futures existed. One such product is an alpha fund. The idea behind such a fund is to capture the stock selection ability of a set of analysts without being subject to market risks. The implementation of the concept simply involves selling enough futures on the S&P index so that the sum of the betas on the futures and the fund's stock portfolio equals zero. Thus the fund has a beta of zero. Assume that the stock portion of the fund has a beta of one. Then the amount of futures to be written equals the value of the fund. From an earlier section we know that the futures price is the spot price adjusted up by the risk-free rate and down by the dividend rate. If the stock portfolio has the same dividend rate as the S&P index, its dividend rate will match the minus dividend term in the equation valuing futures. Thus the return on the fund should be equal to the return on T-bills plus any alpha or greater than equilibrium return earned on its stock portfolio. Such funds are called alpha funds.

A second set of products involving futures stems from the fact that futures can be used to (almost) replicate puts and calls as well as to replicate stocks and bonds. In the section of this chapter on pricing, we have shown how futures are priced by their ability to replicate existing financial instruments. Holding T-bills and buying futures on the S&P index is almost the same as investing in the S&P index.[1] Similarly, selling T-bills and selling futures on the S&P index is almost the same as shorting the S&P index.

We showed in previous chapters that puts and calls can be replicated by dynamically changing the mix of instruments in a portfolio. For example, a put on a stock could be replicated by buying T-bills and shorting the stock. But now we know that shorting the stock can be replicated by borrowing and selling futures on the stock. Thus the put or call may be replicated by using futures in combination with lending and borrowing. This is particularly important in artificial puts (portfolio insurance) constructed through dynamic portfolio rebalancing. This rebalancing can take place at a much lower cost using futures than it can through changing the bond–stock mix. This has led to the creation of products using futures that attempt to replicate holding:

1. T-bills plus calls on long-term bonds
2. T-bills plus calls on stocks
3. Long-term bonds or bond portfolios plus puts on these bonds
4. Stocks or stock portfolios plus puts on these stocks

In closing we should mention that just as a position in futures plus a position in the underlying instrument plus T-bills can be used to replicate calls and/or puts, calls and/or puts can be used in combination with a position in the underlying instruments to replicate futures.

[1]The word *almost* is used because of the effect of marking to the market and the uncertainty of the dividend stream on stocks.

NONFINANCIAL FUTURES AND COMMODITY FUNDS

This chapter is primarily concerned with a discussion of financial futures. Before closing, though, we should mention that there are a tremendous number of nonfinancial commodity futures. Futures exist on a range of additional assets, from those that are thought of as being close to financial assets like silver and gold to those that are almost never thought of as financial futures like hog bellies. In the late 1960s and 1970s, with the tremendous increase in inflation in the American economy, interest grew in both commodity futures and financial futures as hedges against inflation. Actually, returns on commodity futures should reflect only unanticipated inflation as anticipated inflation should already be incorporated in the pricing of the commodity futures. One has to be cautious about interpreting the returns on commodity futures. Commodity futures are used heavily to hedge away the risk faced by producers and manufacturers of products; hence supply, demand, and prices are heavily affected by end product demand and prices. Roll (1985) found that orange juice futures prices were affected by and predictive of weather. In addition, since price is affected by unanticipated inflation rather than inflation itself, to decide on the timing of purchases of futures one has to predict unanticipated inflation. In other words one has to be a better predictor of inflation than the aggregate of investors (the market). Finally, there is the problem of computing a rate of return on a position in futures. Recall that no money changes hands when futures are bought or sold. Only a margin is posted and that can often be posted in the form of T-bills. Very little evidence exists on the rate of return on futures investing. That which does exist makes arbitrary assumptions about the way to compute rates of return and studies a period of time when inflation went from close to zero to over 10%. While it is worthwhile examining these results, one should be somewhat cautious about generalizing from them. Bodie (1983) and Bodie and Rosansky (1980) have studied the performance of an equally weighted portfolio of 23 commodity futures during the period of 1950 through the 1980s. Their data show that over the period the return and risk characteristics of financial futures were very close to that of the S&P 500, but because of a negative correlation between stocks and futures, a portfolio composed of long positions in commodity futures should be included in an investor's optimum portfolio.

There is another approach to the problem of the return on futures and that is to study the performance of publicly traded commodity funds. That is the subject to which we now turn.

Since the 1980s, there has been an explosion of public commodity funds. These funds are similar to mutual funds in the sense that investors buy shares and the proceeds are pooled and managed by a professional manager. Commodity funds invest in financial futures as well as commodity futures. Managers can and will go short as well as long. Thus commodity fund managers are trying actively to guess the course of futures prices rather than taking a passive strategy. Table 24.4 shows the return characteristics of these funds relative to some standard market indexes. The S&P index and the Shearson index can be considered to be, respectively, the return on an index fund of stocks and the return on an index fund of bonds.

Table 24.4 Returns and Risk of Different Investments, 1980–88

Instrument	Average Annual Returns	Standard Deviation of Monthly Returns
Common stocks	14.88%	4.91%
Shearson bond index	11.40%	2.38%
Commodity funds	2.26%	10.4%

Source: Elton, Gruber, and Rentzler (1990).

The striking feature of Table 24.4 is the high variability of the return. The standard deviation of the commodity funds is $2\frac{1}{2}$ to 4 times that of either bond funds or stock funds. The correlation coefficient has been estimated as 0.12 between commodity funds and common equity and -0.03 with the Shearson bond index (see Elton, Gruber, and Rentzler, 1987). Given the low realized returns and the high standard deviation of returns on these funds, evidence in this period indicates that an average commodity fund should not be added to a stock or bond portfolio despite the low correlation. Past return may not be predictive of future returns, but the characteristics shown in Table 24.4 would suggest that commodity funds are not useful additions to bond or stock portfolios.

In a more recent study Bhardwaj et al. (2009) find that for the years 1991–2009 commodity funds have a return which is close to and statistically undistinguished from the return on Treasury bills, but they have a much higher risk.

Despite the evidence against skill in actively managed commodity funds, commodity futures as an asset class have some potentially attractive characteristics. Gorton and Rouwenhorst (2006) constructed an investment index based on rolling long positions of nonfinancial commodity futures. Using a half-century of commodities futures data, they found that the strategy had a risk, return, and Sharpe ratio similar to an equity index, but with low to negative correlation to the stock market. In addition, this passive commodity futures portfolio served as an inflation hedge. This evidence had a major effect on investment managers in the 2000s and stimulated a demand for investment products replicating passive commodity futures investment. Interestingly, Tang and Xiong (2012) found that commodity futures have become more correlated in recent years, possibily as a result of correlated investor demand.

QUESTIONS AND PROBLEMS

1. Given the following data, what is the arbitrage with no transaction costs? What is the size of the transaction costs necessary to negate the arbitrage?

A.	S&P 6-month futures contract	$200
B.	S&P current value	$190
C.	6-month interest rate	6%
D.	Present value of dividends on stocks in S&P index over 6 months	$4

2. Assume that General Mills, a user of wheat, and wheat farmers have the same distributional assumptions about future wheat prices. Does a futures contract make economic sense from both points of view? If yes, why?

3. The spot rate (current rate) for Japanese yen is 120 yen to the dollar, whereas the one-year futures rate is 115. If one-year interest rates in Japan are 4%, what is the implied one-year interest rate in the United States, assuming interest rate parity?

4. Assume you believe that the yield curve will flatten and therefore the spread between long and short rates will narrow. Furthermore, assume others do not share this belief. What action in the futures market should you take to capitalize on your beliefs?

5. Assume you are a bond portfolio manager with $100 million of 20-year corporates. Further assume you wish to hold one-year corporates. Assuming for the moment the availability of any future you wish, design a strategy using futures to accomplish this switch. How would this be accomplished using futures that are traded? What is the additional risk?

6. As a treasurer of the company, you wish to issue $40 million of 10-year bonds. You believe it will take three months before the issue can be floated and that interest rates will rise. You wish to lock in today's rates. Discuss how this can be done using futures contracts.

BIBLIOGRAPHY

1. Adler, Michael, and Detemple, Jerome B. "On the Optimal Hedge of a Non-traded Cash Position," *Journal of Finance*, **43,** No. 1 (March 1988), pp. 143–153.
2. Aggarwal, Raj, and Sundaraghavan, P. S. "Efficiency of the Silver Futures Market," *Journal of Business Finance*, **11,** No. 1 (March 1987), pp. 49–64.
3. Anderson, Ronald W., and Danthine, Jean-Pierre. "Hedging and Joint Production: Theory and Illustrations," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 487–497.
4. Bernard, Victor L., and Frecka, Thomas J. "Commodity Contracts and Common Stocks as Hedges against Relative Consumer Price Risk," *Journal of Financial and Quantitative Analysis*, **22,** No. 2 (June 1987), pp. 169–188.
5. Bhardwaj, Getesh, Goton, Gary, and Rouwenhorst, K. Geert. "Fooling Some of the People All of the Time: The Difficult Performance and Persistence of Community Trading Advisors." Yale ICF working paper No. 08-21 (2008).
6. Black, Fischer. "The Pricing of Commodity Contracts," *Journal of Financial Economics*, **3,** No. 1 (Jan.–March 1976), pp. 167–179.
7. Bodie, Zvi. "Commodity Futures as a Hedge against Inflation," *Journal of Portfolio Management*, **9,** No. 3 (Spring 1983), pp. 12–17.
8. Bodie, Zvi, and Rosansky, Victor. "Risk and Return in Commodity Futures," *Financial Analysts Journal*, **36,** No. 3 (May 1980).
9. Breeden, Douglas T. "Consumption Risk in Futures Markets," *Journal of Finance*, **35,** No. 2 (May 1980), pp. 503–520.
10. Capozza, Dennis, and Cornell, Bradford. "Treasury Bill Pricing in the Spot and Futures Markets," *Review of Economics and Statistics*, **61,** No. 9 (Nov. 1979).
11. Chang, Eric C. "Returns on Speculators and the Theory of Normal Backwardation," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 193–208.
12. Cornell, Bradford. "Taxes and the Pricing of Treasury Bill Futures Contracts: A Note," *Journal of Finance*, **36,** No. 5 (Dec. 1981), pp. 1169–1176.
13. Cornell, Bradford, and French, Kenneth R. "Taxes and the Pricing of Stock Index Futures," *Journal of Finance*, **38,** No. 3 (June 1983), pp. 675–694.
14. Cornell, Bradford, and Reinganum, Marc R. "Forward and Future Prices: Evidence from the Foreign Exchange Markets," *Journal of Finance*, **36,** No. 5 (Dec. 1981), pp. 1035–1046.
15. Cox, John C., Ingersoll, Jonathan E., and Ross, Stephen A. "The Relation between Forward Prices and Future Prices," *Journal of Financial Economics*, **9,** No. 4 (1981), pp. 321–346.
16. Dusak, Katherine. "Futures Trading and Investor Returns: An Investigation of Commodity Market Risk Premiums," *Journal of Political Economy*, **81,** No. 6 (Nov.–Dec. 1973), pp. 1306–1387.
17. Ederington, Louis. "The Hedging Performance of the New Futures Market," *Journal of Finance*, **34,** No. 1 (Mar. 1979), pp. 157–170.
18. Elton, Edwin, Gruber, Martin, and Rentzler, Joel C. "Intra-day Tests of the Efficiency of the Treasury Bill Futures Market," *Review of Economics and Statistics*, **66,** No. 1 (Feb. 1984), pp. 129–137.
19. ——. "Professionally Managed, Publicly Traded Commodity Funds," *Journal of Business*, **60,** No. 2 (April 1987), pp. 175–199.
20. ——. "New Public Offerings Information, and Investor Rationality: The Case of Publicly Offered Commodity Funds," *Journal of Business*, **62,** No. 1 (Jan. 1989), pp. 1–15.
21. ——. "Publicly Offered Commodity Funds," *Financial Analyst Journal*, **46,** No. 1 (July–Aug. 1990).
22. Fama, Eugene. "Forward Rates as Predictors of Future Spot Rates," *Journal of Financial Economics*, **3,** No. 4 (Oct. 1976), pp. 361–377.
23. Figlewski, Stephen. "Futures Trading and Volatility in the GNMA Market," *Journal of Finance*, **36,** No. 2 (May 1981).
24. ——. "Hedging Performance and Basis Risk in Stock Index Futures," *Journal of Finance*, **39,** No. 3 (July 1984), pp. 657–669.

25. Forsythe, Robert, Palfrey, Thomas R., and Plott, Charles R. "Futures Markets and Informational Efficiency: A Laboratory Examination," *Journal of Finance*, **39,** No. 4 (Sept. 1984), pp. 955–982.
26. Gay, Gerard D., and Manaster, Stephen. "Hedging against Commodity Price Inflation: Stocks and Bills as Substitutes for Futures Contracts," *Journal of Business*, **55,** No. 3 (July 1982), pp. 317–344.
27. ——. "Hedging against Commodity Price Inflation. Stocks and Bills as Substitutes for Futures Contracts," *Journal of Business*, **55,** No. 3 (July 1983), pp. 317–343.
28. ——. "The Quality Option Implicit in Futures Contracts," *Journal of Financial Economics*, **13,** No. 3 (Sept. 1984), pp. 353–370.
29. Gorton, G., and Rouwenhorst, G. "KG Facts and Fantasies about Commodity Futures," *Financial Analysts Journal*, **6,** No. 2 (2006), pp. 47–68.
30. Hartzmark, Michael L. "Returns to Individual Traders of Futures: Aggregate Results," *Journal of Political Economy*, **95,** No. 6 (Dec. 1987), pp. 1292–1306.
31. Hilliard, Jimmy E. "Hedging Interest Rate Risk with Futures Portfolios under Term Structure Effects," *Journal of Finance*, **39,** No. 5 (Dec. 1984), pp. 1547–1570.
32. Ho, Thomas S. Y. "Intertemporal Commodity Futures Hedging and the Production Decision," *Journal of Finance*, **39,** No. 2 (June 1984), pp. 351–376.
33. Hsieh, David A., and Kulatilaka, Nalin. "Rational Expectations and Risk Premia in Forward Markets: Primary Metals at the London Metals Exchange," *Journal of Finance*, **37,** No. 5 (Dec. 1982), pp. 1199–1208.
34. Jacobs, Rodney L. "The Effect of Errors in Variables on Tests for a Risk Premium in Forward Exchange Rates," *Journal of Finance*, **37,** No. 3 (June 1982), pp. 667–678.
35. Jagannathan, Ravi. "An Investigation of Commodity Futures Prices Using the Consumption-Based Intertemporal Capital Asset Pricing Model," *Journal of Finance*, **40,** No. 1 (March 1985), pp. 175–192.
36. Jarrow, Robert A., and Oldfield, George S. "Forward Contracts and Futures Contracts," *Journal of Financial Economics*, **9,** No. 4 (Dec. 1981), pp. 373–382.
37. Kamara, Avraham, and Siegel, Andrew F. "Optimal Hedging in Futures Markets with Multiple Delivery Specifications," *Journal of Finance*, **42,** No. 4 (Sept. 1987), pp. 1007–1021.
38. Kilcollin, Thomas Eric. "Difference Systems in Financial Futures Markets," *Journal of Finance*, **37,** No. 5 (Dec. 1982), pp. 1183–1198.
39. Park, Soo-Bin. "Spot and Forward Rates in the Canadian Treasury Bill Market," *Journal of Financial Economics*, **10,** No. 1 (March 1982), pp. 107–114.
40. Rendleman, Richard, and Carabini, Christopher. "The Efficiency of the Treasury Bill Futures," *Journal of Finance*, **34,** No. 4 (Sept. 1979), pp. 895–914.
41. Richard, Scott F., and Sundaresan, M. "A Continuous Time Equilibrium Model of Forward Prices and Futures Prices in a Multigood Economy," *Journal of Financial Economics*, **9,** No. 4 (Dec. 1981), pp. 347–372.
42. Roll, Richard. "Orange Juice and Weather," *American Economic Review*, **74,** No. 5 (Dec. 1985), pp. 861–881.
43. Tang, Ke, and Wei, Xiong. "Index Investment and the Financialigation of Commodities," *Financial Analysts Journal*, **68,** No. 6 (2012), pp. 54–74.
44. Williams, Jeffrey. "Futures Markets: A Consequence of Risk-Aversion or Transactions Costs," *Journal of Political Economy*, **95,** No. 5 (Oct. 1987), pp. 1000–1023.

Part 5

EVALUATING THE INVESTMENT PROCESS

25

Mutual Funds

Mutual funds have existed for over 200 years. The first mutual fund was started in Holland in 1774, but the first mutual fund did not appear in the United States for 50 years, until 1824. Since then the industry has grown in size to $24 trillion worldwide and over $11.6 trillion in the United States. The importance of mutual funds to the U.S. economy can be seen by several simple metrics:[1]

1. Mutual funds in terms of assets under management are one of the two largest financial intermediaries in the United States.
2. Almost 50% of American families own mutual funds.
3. Over 50% of the assets of defined contribution pension plans and individual retirement plans are invested in mutual funds.

In the United States, mutual funds are governed by the Investment Company Act of 1940. Under law, mutual funds are legal entities that have no employees and are governed by a board of directors (or trustees) who are elected by the fund investors. Directors outsource all activities of the fund and are charged with acting in the best interests of the fund investors.

Mutual funds tend to exist as members of fund complexes or fund families. There are 16,506 funds in the United States. Of these, 8,684 are open-end funds, which are distributed by 713 fund families.[2] Funds differ from each other by the type of securities they hold, the services they provide, and the fees they charge. The sheer number of funds makes evaluation of performance important. Data, transparency, and analysis become important in selecting funds.

Usually when people talk about mutual funds, they are referring to open-end mutual funds, but there are three other types of mutual funds: closed-end funds, exchange-traded funds, and unit investment trusts. The size of each type of mutual fund, both in assets under management and number of funds, is presented in Tables 25.1 and 25.2. We will refer to these data throughout the introduction. Examining each type as a percentage of the total assets in the

[1]All descriptive statistics in this section as of the end of 2011 (or the last available data on that date) unless otherwise noted.

[2]The assets in fund families are highly concentrated, with the 10 largest families managing 53% of the assets in the industry and the top 25 families managing 73%. The number of mutual funds reported excludes 6,022 Unit Investment Trusts. All numbers come from Investment Company Institute data.

Table 25.1 Total Net Assets by Type

Billions of Dollars, Year-End, 1995–2010					
	Open-End Mutual Funds[a]	Closed-End Funds	ETFs[b]	UITs	Total[c]
1995	$2,811	$143	$1	$73	$3,028
1996	3,526	147	2	72	3,737
1997	4,468	152	7	85	4,712
1008	5,525	156	16	94	5,791
1999	6,846	147	34	92	7,119
2000	6,965	143	66	74	7,248
2001	6,975	141	83	49	7,248
2002	6,383	159	102	36	6,680
2003	7,402	214	151	36	7,803
2004	8,095	254	228	37	8,614
2005	8,891	277	301	41	9,510
2006	10,398	298	423	50	11,168
2007	12,002	313	608	53	12,977
2008	9,604	186	531	29	10,349
2009	11,120	225	777	38	12,161
2010	11,821	241	992	51	13,104
2011	11,621	239	1,048	60	12,968

[a]Mutual fund data include only mutual funds that report statistical information to the Investment Company Institute. The data do not include mutual funds that invest primarily in other mutual funds.
[b]ETF data prior to 2001 were provided by Strategic Insight Simfund. ETF data include investment companies not registered under the Investment Company Act of 1940 and exclude ETFs that invest primarily in other ETFs.
[c]Total investment company assets include mutual fund holdings of closed-end funds and ETFs.

Note: Components may not add to the total because of rounding.
Sources: Investment Company Institute and Strategic Insight Simfund

industry, we find at the end of 2011 that open-end mutual funds are 89.6%, closed-end funds 1.8%, exchange-traded funds 8.1%, and unit investment trusts less than 0.4%.

The breakdown by number of funds has very different percentages: 53%, 3.8%, 7.1%, and 36.5% for open end, closed-end, ETFs, and Unit Investment Trusts (UITs), respectively. While there are a very large number of UITs, they constitute a small percentage of the assets under management by mutual funds.

In this chapter we discuss the three largest types of funds, with emphasis on the unique aspects of each. Although we discuss each type of fund in this chapter, the discussion of open-end funds is less detailed as Chapter 26 is devoted to an explanation of the performance of open-end funds.

OPEN-END MUTUAL FUNDS

In terms of number of funds and assets under management, open-end mutual funds are by far the most important form of mutual funds. What distinguishes them from other forms is that the funds can be bought and sold anytime during the day, but the price of the transaction is set at the net asset value of a share at the end of the trading day, usually 4 PM. It is both the ability to buy and sell at a price (net asset value) which will be determined after the buy or sell decision, and the fact that the other side of a buy or sell is the fund itself, that differentiates this type of fund from other types.

Table 25.2 Number of Mutual Funds by Type

	Year-End, 1995–2010				
	Open-End Mutual Funds[a]	Closed-End Funds	ETFs[b]	UITs	Total[c]
1995	5,761	500	2	12,979	19,242
1996	6,291	497	19	11,764	18,573
1997	6,778	487	19	11,593	18,877
1008	7,489	492	29	10,966	18,976
1999	8,003	512	30	10,.414	18,959
2000	8,370	482	80	10,072	19,004
2001	8,518	492	102	9,295	18,407
2002	8,511	545	113	8,303	17,472
2003	8,426	584	119	7,233	16,362
2004	8,415	619	152	6,499	15,685
2005	8,449	635	204	6,019	15,307
2006	8,721	647	359	5,907	15,634
2007	8,747	664	629	6,030	16,070
2008	8,884	643	743	5,984	16,254
2009	8,617	628	820	6,049	16,114
2010	8,545	624	950	5,971	16,090
2011	8,684	634	1,166	6,022	16,506

[a]Investment company data include only investment companies that report statistical information to the Investment Company Institute.
[b]The data include mutual funds that invest primarily in other mutual funds.
[c]ETF data prior to 2001 were provided by Strategic Insight Simfund. ETF data include investment companies not registered under the Investment Company Act of 1960 and ETFs that invest primarily in other ETFs.

Sources: Investment Company Institute and Strategic Insight Simfund

Mutual funds are subject to a single set of tax rules. To avoid taxes, mutual funds must distribute by December 31 98% of all ordinary income earned during the calendar year and 98% of all realized net capital gains earned during the previous 12 months ending October 31. They rarely choose not to do so. They can lower their capital gains distributions by offsetting gains with losses and by occasionally paying large investors with a distribution of securities rather than cash.[3]

Open-end funds are the mutual fund type that has by far the largest amount of assets under management, and they have had phenomenal growth in assets. Starting with a 1995 value of $2.8 trillion, assets under management of open-end funds grew to $11.6 trillion by the end of 2011, a growth rate of more than 9% per year. The growth rate of open-end funds is only exceeded by the growth rate of ETFs, which started at a base of only $1 billion in 1995. The tremendous growth in assets under management of open-end funds was fueled by two sources: a high rate of return in the capital markets and the huge inflows of new capital due in large part to the growth in the private pension market in the United States. The importance of the private pension market in the United States to mutual funds in the United States can be seen by the fact that by the end of 2011, there were $4.68 trillion invested in mutual funds by private pensions, which represents 36% of the assets held by mutual funds.

Since 1995 net inflows and the return earned on these net inflows have accounted for about half of the increase in the assets of open-end mutual funds. In this period, net inflows have averaged about $190 billion a year and have been positive in every year, except for

[3]These tax rules apply to all types of mutual funds, not just open-end funds.

2008 and 2011. As a percentage of beginning assets, net inflows to open-end mutual funds have been about 3.3% per year, with a high of 8.3% in 1995 and a low of −1.9% in 2008. Other than 2008 and 2011, the lowest net inflow was 1.79% in 2001.

It is worthwhile reviewing the history of the size and growth of the major types of open-end funds. As of the end of 2011, the breakdown by type of open-end fund was 45% equity funds, 7% hybrid funds, 25% bond funds, and 23% money market funds. This was not always the case. There was very little growth in assets of the mutual fund industry in the early to mid-1970s. In the late 1970s and early 1980s most of the growth in assets was in money market funds. Growth was spurred by the granting of the right for money market funds to have check-writing privileges. By 1981, money market funds contributed 77% of the assets in the industry. The next segment of the industry to grow was bond funds. Investors learned that they could earn higher rates of interest by buying bond funds rather than by buying money market funds. The last sector to grow in terms of assets under management was stock funds. It was not until 1993 that assets in stock funds exceeded both bond funds and money market funds in size. Since then, stock funds have remained the largest sector of the industry, only being temporarily exceeded in size by one other type (money market funds) during the market decline in 2008.

As will become clear when we discuss performance, the size of the expenses ratio plays a major role in the measured performance of the mutual fund industry and the relative performance of individual mutual funds. Expenses usually consist of two parts: an annual fee captured in the fund's expense ratio and a one-time fee called the front-end or back-end load fee. Fees in the mutual fund industry have decreased markedly in the past 20 years. We will use a metric to measure fees developed by the Investment Company Institute which adds to the funds' annual expense ratio an estimate of the annualized cost an investor potentially pays due to a one-time sales load. They arrive at an overall fee level by weighting the fees on any fund by the assets under management of that fund. As shown in Table 25.3, average fees are higher for stock funds than for bond funds, but both have decreased by 50% since 1990.

There are several reasons for this decrease. First is the increased importance of passively managed index funds, which typically have very low expense ratios. Today, more than 13% of the assets held by mutual funds are held in passive portfolios. Second, more and more ownership of mutual funds occurs through employee-sponsored retirement plans; load fees are often much lower or nonexistent for these plans, and expense ratios are also generally lower. A third factor is the increased sensitivity of investors to expenses, causing a reduction in load fees and yearly expenses. Between 2000 and 2011, 75% of net new cash flow went to the 25% of the funds with the lowest expense ratio.

Another factor leading to lower expenses is the growth of mutual fund size and individual account size in the industry. Both of these factors result in economies of scale, which result in lower costs, at least some of which are passed on to investors.

Before leaving this section, it is interesting to place the U.S. industry in the context of the world market for mutual funds. As shown in Table 25.4, in 2010 the worldwide market for

Table 25.3 Expense Ratio in Annual Percentage[a]

	1990	2000	2010	2011
Stock funds	2.00	1.28	0.95	0.93
Bond funds	1.85	1.00	0.72	0.66

[a]ICI estimate of expenses is lower than often reported. This is because they weight expense ratios by the size of the fund and large funds tend to have larger expense ratios than small funds. This is often attenuated by the fact that many of the large funds are index funds. If fees were computed by averaging the fees treating each fund equally, the average fee for equity funds would be 143 basis points in 2011.

Table 25.4 Total Net Assets of Mutual Funds (in Billions)

	2004	2006	2008	2010	2011
World	16,153	21,808	18,920	24,699	23,800
Americas	8,781	11,470	10,582	13,586	13,513
U.S.	8,095	10,398	9,604	11,821	11,621
Europe	5,640	7,804	6,231	7,903	7,720
Asia and Pacific	1,678	2,456	2,036	3,067	2,921
Africa	54	78	69	142	125

open-end mutual funds was $23.8 trillion, while the mutual fund assets in the United States were $11.6 trillion.

It is clear from these numbers that the U.S. mutual fund industry has the most assets under management but that the growth rate of assets under management in the rest of the world has exceeded the growth rate in the United States. Because of the size and importance of open-end mutual funds, we devote Chapter 26 to their performance and characteristics.

CLOSED-END MUTUAL FUNDS

Closed-end mutual funds, like open-end mutual funds, hold securities as their assets and allow investors to buy and sell shares in the fund. The difference is that shares in a closed-end fund are traded on an exchange and have a price determined by supply and demand, which (unlike open-end funds) can, and usually does, differ from the net asset value of the assets of the fund. Furthermore, shares can be bought or sold at any time the market is open at the prevailing market price, while open-end funds are priced only once a day. Perhaps the easiest way to think of closed-end funds is a company that owns securities rather than machines. The difference between the price at which a closed-end fund sells and its net asset value has been the subject of a large amount of analysis and will be reviewed in great detail later in this chapter. We will simply note here that closed-end stock funds tend to sell at a discount from the net asset value of their holdings.

The composition of the $239 billion in closed-end funds is different from the composition of open-end funds. Bond funds constitute 61% of the assets in closed-end funds and stock funds 39% of the assets. If we restrict the analysis to funds holding domestic assets, the percentages are 69% to bonds and 31% to equity. This stands in contrast to open-end funds, where the reverse is true. Equity funds hold a much larger percentage of the assets.

While there have been a huge number of interesting articles discussing closed-end funds and the anomalies they present, we have decided to limit the discussion to two subjects: the discount or premium at which closed-end funds sell and the reasons for the existence of closed-end funds.[4]

Explaining the Discount

Explanations for the discount (the amount by which the value of the holdings of a fund exceed the fund's market value) at which closed-end funds sell include liquidity of investments, management fees, management ability, tax liabilities, sentiment, greater risk of

[4]The most cogent discussion of the major anomalies in the pricing of the closed-end funds is presented by Lee, Shleifer, and Thaler (1990). These include the premium for new funds, the cross-sectional and intertemporal behavior of discounts, and the price behavior when funds are terminated.

closed-end fund returns compared to returns on their assets, and uncertainty about the size of future discounts.

When one buys a closed-end fund that holds securities with a capital gain, one owns a share in the assets and a share in a future potential tax liability. However, given the high turnover of most domestic closed-end funds, the tax overhang should be small.[5] Malkiel estimates that even with very high estimates of capital gains overhang, that overhang can account for only a small part of the discount at which closed-end funds sell. In addition, as Lee, Shleifer, and Thaler (1991) point out, a capital gain explanation for discounts predicts that discounts should increase when returns are high, but in fact there is no correlation between discounts and returns. However, in more recent articles, Brennan and Jain (2007) examine the behavior of closed-end funds around capital gains and dividend distributions and find evidence that there is an effect of tax overhang. There is no doubt that tax overhang affects the pricing of closed-end funds. However, it appears to account for only a small portion of the discount.

A number of authors have investigated explanations for the discount using expenses or the trade-off between management ability and expenses. Kumar and Noronha (1992) find a positive relationship between expenses and discounts. Expenses should be examined in combination with performance. If management produces superior performance before expenses, the question remains whether the net result of management ability and expenses can account for the discount. Cherkes et al. (2009) argue that the discount can be explained by the capitalized value of the services management adds less the capitalized value of the cost of such service.

Berk and Stanton (2007) provide one of the more compelling explanations of the discount. Their argument is that if management is entrenched, poor management relative to expenses leads to a discount. However, if management is free to leave when performance is good, management will capture the extra performance in higher fees or leave for a different job. Thus the balance of expenses and performance means that an average fund sells at a discount.

Two additional plausible explanations have been offered for the size and existence of the discount: one based on behavioral and one based on capital market characteristics. A well-known series of papers by Lee, Shleifer, and Thaler (1990, 1991), DeLong and Shleifer (1992), and Chopra, Shleifer, and Thaler (1993) explains the discount on closed-end funds by the irrational sentiment of retail investors.

LS&T (1990) hypothesized that retail investors are at times overly optimistic and at other times overly pessimistic. They argue that closed-end funds tend to be held by retail investors and that the added risk introduced by irrational retail investors means that closed-end funds sell at a discount. Irrational sentiment risk then becomes a systematic influence that affects not just closed-end funds but any investment (e.g., small stocks) held by retail as opposed to institutional investors.

Elton, Gruber, and Busse (1996) offered an alternative explanation for the discount on closed-end domestic stock funds based on the market characteristics of these funds. They show that the loadings (betas) on the Fama–French systematic factors, the market, the small versus large stock index, and a value minus growth index, are higher for the return on closed-end funds than they are for the returns on the securities these funds hold. Why do these differences in sensitivities arise?

Elton, Gruber, and Busse (1996) found the average market value of stocks held by closed-end stock funds was $5,572 million, while the average market value of the funds holding these stocks was $343 million. Similarly, the average market-to-book ratio of the stocks

[5]Tax overhang is only a deterrent to the extent the capital gains are realized while the investor is in the fund. Tax overhang could be correlated with turnover or performance, which might impact the results.

held by funds was 3.9, while for the fund itself, it was 0.9. This explains why the loadings on two recognized risk factors (small-large and value-growth) were so much larger for the funds than on the portfolio of securities they held. The higher loadings and positive factor prices mean more risk for the closed-end funds than the portfolio they hold. The higher risk must be compensated for by higher expected return. The only way this can happen is for the average price for closed-end funds to be lower than the NAV on these funds.

Either of the explanations (irrational sentiment as a systematic influence or the Fama–French model combined with the different risks associated with the fund and the portfolio they hold) can be used to explain the persistent discount for closed-end stock funds and, to a large extent, the movement of the discount over time.

Why Closed-End Funds Exist

There is a second topic of great interest with respect to closed-end funds: why do they exist at all? The classic reason given for the existence of closed-end funds is that their organizational form allows them to hold fewer liquid assets and to hold less cash. This reason has been explored both theoretically and empirically in a series of papers, perhaps most cogently in Cherkes, Sagi, and Stanton (2009) and Deli and Varma (2002). Because closed-end funds are not subject to inflows when investors buy a fund or to key importance outflows of cash when investors choose to sell a fund, they argue that closed-end funds can hold more illiquid assets and less cash than open-end funds. This is, no doubt, an explanation for the creation of many types of closed-end funds. Cherkes et al. do a thorough job of exploring a liquidity-based theory of closed-end funds. Deli and Varma test and find evidence that closed-end funds are more likely to hold securities in illiquid markets.

While the advantage of organization structure which allows for holding illiquid assets can account for some of the popularity of closed-end funds, there is another advantage of organizational structure that has not received as much attention. Closed-end funds, unlike open-end funds, have the ability to use large amounts of leverage to finance their investments.

Elton, Gruber, and Blake (2013) design a study to more clearly show the impact of leverage. They study closed-end bond funds because there are many more closed-end bond funds than closed-end stock funds. Furthermore, there are a number of closed-end bond funds, each of which can be matched with an open-end bond fund with the same portfolio manager and same objectives, and which are sponsored by the same fund family. By studying matched pairs of funds, the effects of many of the influences affecting performance can be held constant. EG&B show that the characteristics of the assets and the returns on the assets earned by the open- and closed-end funds in the matched sample are almost identical. The difference between the open- and closed-end funds is the increased return to investors due to the use of leverage: leverage ratios for the closed-end funds averaged more than 50%. Leverage is advantageous to closed-end funds because they borrow short term, usually in the form of floating rate preferred stock, and invest in longer-term bond funds. The advantage of fund leverage rather than investor leverage arises from at least three factors: (1) tax law (interest paid on the preferred stocks issued by municipal closed-end bond funds is not taxable to the holder of the preferred stock, and for nonmunicipal closed-end funds, the interest paid on preferred stock is taxed at a lower rate than the interest paid on debt instruments), (2) limited liability to the holder of fund shares, and (3) lower borrowing costs to the fund compared to investor borrowing costs. For example, the borrowing rate paid on preferred stock by municipal closed-end bond funds is considerably lower than the federal fund rate.[6]

[6]The fact that the vast majority of closed-end bond funds tend to employ leverage ratios close to the maximum allowed by law is evidence that managers of these funds believe that leverage is important.

The research proceeds to show that the leveraged closed-end bond funds are a more desirable asset to add to a portfolio of stocks or bonds than unlevered closed-end funds or open-end funds. Furthermore, in a larger sample of closed-end bond funds, differences in leverage account for more than 24% of the cross-sectional differences in discount, and discounts vary over time as a function of the difference between long rates and short rates, a measure of the desirability of leverage.

EXCHANGE-TRADED FUNDS (ETFS)

Exchange-traded funds are a recent phenomenon, with the first fund (designed to duplicate the S&P 500 index) starting in 1993. They are a fast growing segment of the mutual fund industry. They are very much like closed-end funds, with one exception. Like closed-end funds, they trade at a price determined by supply and demand and can be bought and sold at that price during the day. They differ in that at the close of the trading day, investors can create more shares of ETFs by turning in a basket of securities which replicate the holdings of the ETF or can turn in ETF shares for a basket of the underlying securities. This eliminates one of the major disadvantages of closed-end funds, the potential for large discounts. If the price of an ETF strays very far from its net asset value, arbitrageurs will create or destroy shares, driving the price very close to the net asset value. The liquidity which this provides to the market, together with the elimination of the risk of large deviations of price from net asset value, has helped account for the popularity of ETFs.

The composition of exchange-traded funds is very different from the composition of other types of mutual funds. The biggest difference is in the importance of index funds to this part of the industry. In 2011, 90% of the assets in exchange-traded funds were held in passive funds. This contrasts with the 13% held in open-end funds. As late as 2007, there were virtually no actively managed exchange-traded funds.

Exchange-traded funds have been organized under three different sets of rules. The differences in organizational structure are important because they can affect what actions the ETF can take in managing the portfolio. The original ETF (spider) was organized as a trust. The trust structure requires exact replication of the index (rather than sampling). Furthermore, it does not allow security lending or the use of futures and requires that dividends received from the securities the fund holds be placed in a non-interest-bearing account until they can be disbursed to shareholders. Most ETFs organized after the spiders were organized as managed funds. Managed funds have much greater flexibility, allowing sampling, the purchase and sale of futures, security lending, and the immediate reinvestment of dividends. The third possible organizational structure is a granter trust. Investors in granter trusts hold the shares directly, retaining their voting rights and receiving dividends and spinoffs directly. They can unbundle the trust, selling off some of the companies in the trust. There is no separate management fee; rather, there is a custodian fee for holding the shares. ETFs called “holders” are granter trusts.

ETFs are stocks and trade on exchanges like other stocks. ETFs’ assets are a basket of securities rather than physical assets, and as such they are similar to closed-end funds. They differ from closed-end funds in that new shares can be created or old shares can be deleted every day. For example, the largest ETF is the spider. The spider attempts to mimic the S&P 500 index with one share equal to approximately one-tenth of the price of the S&P 500 index. New or old shares are deleted or created in minimum orders of 50,000 shares for a payment of $3,000, regardless of the number of units involved. At the end of the day the fund posts its holdings (including cash). An investor wishing to create shares turns in a bundle of stock holdings that match the S&P 500 index plus the appropriate amount of cash. There is more creation than deletion, and both are in large amounts.

Creations and deletions occurred on approximately 15% of the trading days. On these days Elton, Gruber, Comer, and Li (2002) report that creations and deletions average over $100 million.

The system of creation and deletion and the ability to arbitrage price and NAV differences means that the price of a share in an exchange-traded fund has historically been close to NAV, unlike the price of closed-end funds. Most exchange-traded funds attempt to match an index and are passive in their investment strategy. The principal issues are as follows:

1. tracking error
2. the relationship of price to NAV
3. their performance relative to other indexing vehicles
4. their use in price formation
5. the effect of leverage
6. active ETFs

Each of these will be discussed in turn.

Tracking error

Tracking error is the performance of the portfolio compared to the performance of the index. For large, well-diversified portfolios like those matching the S&P 500 index, tracking error is minimal and not very important. Using a sample of S&P index funds, Elton, Gruber, and Busse (2004) report average R^2 in excess of 0.9999 for S&P 500 index funds, which means the tracking error for S&P 500 index funds is less than 0.0001%. This should also be true for exchange-traded funds tracking the S&P 500 index.

However, many ETFs attempt to match indexes with sampling techniques rather than replication that exactly matches the index. Likewise, many ETFs attempt to match a country or sector index where a single security represents a large portion of the market and exact replication is not possible because of rules prohibiting more than 5% of the portfolio being invested in a single security. These ETFs can have a serious problem in index replication.

The Relationships of Price to NAV

The process of creation and deletion keeps price and NAV fairly close, particularly at the end of the day. However, there are deviations, and this is a potential cost to an investor who wishes to buy or sell and finds the price differs from NAV in an adverse way. It can, of course, also be a benefit if the investor buys when the price is below NAV and sells when it's above. For actively traded ETFs, prices and NAVs are very close, and differences are transient. Engle and Sarkar (2002) examine differences for actively traded ETFs. The standard deviation of the premium and discounts was around 15 basis points and was less than the bid–ask spread. For less actively traded funds (they used international funds to represent less actively traded funds), the standard deviation is much larger, and deviations can persist over several days.[7]

Performance Relative to Other Instruments

Passive ETFs often match the same index as an open-end index fund. Also, there are sometimes futures on the index that can be used in conjunction with a bond to create a portfolio that matches the index.

[7]Chery (2004) also studies this phenomenon.

How does the performance of these instruments compare? We will compare ETFs and index funds.[8] The difference in performance depends on the skill in matching the index, expenses, charter restrictions, and tax considerations. Even for passive funds that construct their portfolio by exact replication, there can be differences in skill or differences in the actions allowed by the funds' charters that can affect relative performance. Probably the most important factor is how the fund handles changes in the index being matched. There are often large price changes around the time a security enters or leaves an index. The timing of the portfolio changes for the ETF which may represent management skill, or restrictions on the ETFs imposed by its charter, can affect return. Additional skill factors that affect relative performance include ability to lend securities, dealing with tender offers and mergers, policies involving cash, trading strategies, ability to reinvest dividends, transaction costs, and the ability to use (and skill in using) futures. Depending on how the fund was organized, the ETF may or may not have flexibility on these issues. As pointed out by Elton, Gruber, Comer, and Li (2002), ETFs organized as trusts, such as spiders, must hold the dividends received on underlying securities in a non-interest-bearing account where an index fund will reinvest the dividends or earn interest on them. If the market increases, this is a disadvantage to the ETF.[9] If the market decreases and the index funds reinvest dividends, then this is an advantage. Partly because of the disadvantage of holding dividends in a non-interest-bearing account and restrictions on lending, the use of futures and restrictions on rebalancing most ETFs issued after spiders choose a different organizational form. In addition to management skill affecting performance, expenses are a cost to investors and ceteris paribas hurt performance.[10]

The final difference affecting performance is tax considerations. ETFs are considered tax efficient because they generally distribute fewer capital gains than index funds. Capital gains are generated when shares are sold and the price at which they were bought is less than the selling price. The way ETF shares are created and deleted provides ETFs with a chance to maintain a high cost basis on shares in their portfolios. When ETF shares are redeemed, the trustee delivers in-kind securities that comprise the index. The trustee always delivers the lowest-cost shares, keeping the cost basis high. The IRS has ruled that the process of deletion is not a taxable exchange. Thus, if an investor turns in ETFs worth $100 million and the trustee gives the investor securities with a cost basis of $50 million, there are no capital gains taxes on the arbitrageur or the ETF. Poterba and Shoven (2002) studied the capital gain payment on the Vanguard S&P index fund and the ETF spider and found tax considerations gave the spider a tax advantage, but this was not nearly enough to overcome the other considerations that favored the index fund.

Their Use of Price Formation

Hasbrouck (2003) and Schlusche (2009) examine the process of information incorporation when multiple contacts exist on the same index. For example, the S&P 500 index has the spider, an ETF, a floor-traded futures contract, and a small-denomination electronically traded futures contact. Hasbrouck finds in this market information is first incorporated in the small-denomination futures contract. In other markets the results can differ. For example, in the market for the S&P 400 mid-cap, which has an ETF and a futures contract, Hasbrouck (2003) finds information is reflected equally.

---

[8]See Elton, Gruber, Comer, and Li (2002) for a comparison with futures.

[9]Elton, Gruber, Comer, and Li (2002) show how the shortfall of the spider compared to an index fund is a function of market movement and the size of the dividends.

[10]In comparing index funds, Elton, Gruber, and Busse (1996) find that future performance is highly predicted by expense ratios. When they regressed the difference in returns between the fund and the index on the prior year's expense rates, they had a slope of –1.09% with an R^2 of 0.788.

The Effect of Leverage

Several hundred ETFs have been developed that are levered, promising multiples of the daily returns on the index either positive or negative.[11] If a standard ETF return pattern can be expressed as 1x, where x is the index's return, then these products are expressed as 2x, 3x, −2x, and −3x. Unlike normal ETFs that hold the underlying securities, these products are constructed using derivatives. This means that the tax efficiency discussed earlier does not hold because realized gains from derivative contracts are taxed at ordinary income tax rates, and creation and deletion are usually in cash, not in kind. Also, these products have much higher expense ratios than standard ETFs. These products are designed for short-term traders. Investors holding them over a long period need not get the promised multiple return (2x or 3x) over the longer period. This occurs as shown because the products are re-levered every day to the stated objective.

The effect of daily re-levering on multiday returns is easy to see with a two-period example. Assume an investor has one dollar, borrows $(m - 1)$ dollars, and invests m dollars in a 1x ETF holding the borrowing at $m - 1$ for both periods. The ending value (ignoring interest on the borrowing and recognizing that $(m - 1)$ is paid back) is

$$m(1 + r_1)(1 + r_2) - (m - 1) \tag{25.1}$$

If the investor invests one dollar in an mx levered ETF, the return is

$$(l + r_1 m)(l + r_2 m) \tag{25.2}$$

For one period the payoff is the same, but because rebalancing occurs, the two-period payoff is different. The difference (return on the levered ETF minus return on "homemade" leverage) is $(m^2 - m)r_1 r_2 > 0$. If $r_1 r_2 > 0$, then the daily rebalancing gives a higher return. If $r_1 r_2 < 0$, then daily rebalancing gives a lower return. Cheng and Madhavan (2009) show that with high volatility and little trend, an investor invested in an mx ETF will get less than mx in return. Given the high fees and that income is mostly ordinary income rather than capital gains, even with an upward trend, an investor is likely to get less than expected over longer time frames. However, an investor may still choose this form of index fund, for it allows a higher level of debt than the investor can get on personal accounts.

Active ETFs

Active ETFs have only recently been introduced and so have not yet been subject to serious academic study. ETFs require daily posting of the portfolio to facilitate creation and deletion. Many trades for mutual funds are executed over several days to mitigate price impacts. Daily reporting of positions can cause front running. This has slowed their introduction.

CONCLUSION

In this chapter we have presented a broad review of the variety and characteristics of mutual funds. In the next chapter we present the major tools for measuring mutual fund performance and evidence on how well the industry has done.

BIBLIOGRAPHY

1. Berk, Jonathan, and Stanton, Richard. "Managerial Ability, Compensation and the Closed End Fund Discount," *Journal of Finance*, **62** (2007), pp. 529–556.

[11]The following analysis is based on Cheng and Madhavan (2009).

2. Brennan, Michael, and Jain, Ravi. "Capital Gains Taxes, Agency Costs and Closed End Fund Discounts," unpublished manuscript, UCLA (2007).
3. Brickley, James, Manaster, Steven, and Schallheim, James. "The Tax Timing Option and the Discounts on Closed End Investments Companies," *Journal of Business*, **64** (1991), pp. 287–312.
4. Chay, J. B., and Tryzinka, Charles. "Managerial Performance and the Cross Sectional Pricing of Closed-End Bond Funds," *Journal of Financial Economics*, **52** (1999), pp. 379–308.
5. Cheng, Minder, and Madhaven, Ananth. "The Dynamics of Levered and Inverse Exchange Traded Funds," *Journal of Investment Management*, (2009), pp. 49–60.
6. Cherkes, Martin. "A Practical Theory of Closed-End Funds as an Investment Vehicle," working paper, Princeton University (2003).
7. Cherkes, Martin, Sagi, Jacob, and Stanton, Richard. "A Liquidity Based Theory of Closed-End Funds," *Review of Financial Studies*, **22** (2009), pp. 257–297.
8. Cherry, Josh. "The Limits of Arbitrage: Evidence from Exchange Traded Funds," unpublished manuscript, University of Michigan (2004).
9. Chopra, Navin, Lee, Charles, Shleifer, Andrei, and Thaler, Richard. "Yes, Discounts on Closed-End Funds are a Sentiment Index," *Journal of Finance*, **48** (1993), pp. 801–808.
10. Deli, Daniel, and Varma, Raj. "Closed-End versus Open: The Choice of Organizational Form," *Journal of Corporate Finance*, **8** (2002), pp. 1–27.
11. Delong, J. Bradford, and Shleifer, Andrei. "Closed-End Fund Discounts: A Yardstick of Small Investor Sentiment," *Journal of Portfolio Management*, **18** (1992), pp. 46–53.
12. Elton, Edwin J., Gruber, Martin J., and Busse, Jeffrey A. "Do Investors Care about Sentiment?" *Journal of Business*, **71** (1996), pp. 475–500.
13. Elton, Edwin J., Gruber, Martin J., Comer, George, and Li, Kai. "Spiders: Where are the Bugs?" *Journal of Business*, **75** (2002), pp. 453–472.
14. Elton, Edwin J., Gruber, Martin J., and Busse, Jeff. "Are Investors Rational? Choices Among Index Funds," *Journal of Finance*, **58** (2004), pp. 427–465.
15. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Why do Closed End Funds Exist? An Additional Explanation for the Growth in Domestic Closed-End Bond Funds" *Journal of Financial and Quantitative Analysis*, **48** (2013).
16. Engle, Robert, and Sarkar, Debojyoti. "Pricing Exchange, Traded Funds," unpublished manuscript, New York University (2002).
17. Fama, Eugene F., and French, Ken R. "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, **51** (1996), pp. 55–87.
18. Farnsworth, Heber, Ferson, Wayne, Jackson, David, and Todd, Steven. "Performance Evaluation with Stochastic Discount Factors," *Journal of Business*, **75** (2000), pp. 473–504.
19. Hasbrouck, Joel. "Intraday Price Formation in the Market for U.S. Equity Markets," *Journal of Finance*, **58** (2003), pp. 2375–2400.
20. Investment Company Factbook. "Investment Company Institute," Washington, DC (2011).
21. Kumar, Raman, and Noronha, G. M. "A Re-examination of the Relationship between Closed-End Fund Discounts and Expenses," *Journal of Financial Research*, **15** (1992), pp. 139–147.
22. Lee, Charles, Shleifer, Andrei, and Thaler, Richard. "Closed End Mutual Funds," *Journal of Economic Perspectives*, **4** (1990), pp. 153–166.
23. Lee, Charles, Shleifer, Andrei, and Thaler, Richard. "Investor Sentiment and the Closed End Fund Puzzle," *Journal of Finance*, **46** (1991), pp. 76–110.
24. Malkiel, Burton. "The Valuation of Closed End Investment Company Shares," *Journal of Finance*, **32** (1977), pp. 847–858.
25. Manzler, David. "Liquidity, Liquidity Risk and the Closed-End Fund Discount," Unpublished manuscript, University of Cincinnati (1990).
26. McConnell, John, and Saretto, Alessio. "Auction Failure and the Market for Auction Rate Securities," *Journal of Financial Economics*, **97** (2010), pp. 451–469.
27. Pontiff, Jeffrey. "Costly Arbitrage, Evidence from Closed-End Funds," *Quarterly Journal of Economics*, **111** (1996), pp. 1135–1151.
28. Pontiff, Jeffrey. "Excess Volatility and Closed-End Funds," *American Economic Review*, **86** (1997), pp. 155–169.

26

Evaluation of Portfolio Performance

An integral part of any decision-making process should be the evaluation of the decision. This is equally true whether investors make their own investment decisions or employ a manager to make them.

A large percentage of investments are made by professional managers. Professionally managed funds include mutual funds, pension funds, college endowments, and discretionary accounts, among others. It is important for an investor utilizing one of these managers not only to evaluate how well the fund has done relative to other funds but also to understand the fund's general policies and to be able to tell how well the fund has followed them. How diversified is the fund? How actively does it try to pursue short-run aberrations in prices? What is the bond–stock mix, and how much does it vary? For the individual investor to understand the risks he is undertaking, the fund's policies and how strictly the manager adheres to them must be known. For the institution that has engaged a professional manager, examining the manager's policies enables the institution to evaluate not only the risks it is undertaking but also the costs of any restrictions it might have placed on the fund manager.

Evaluation is important, not only to the individual or institution who engages a professional money manager but also to the individual who invests personal funds. Once again, evaluation involves more than rating how well the investor has performed compared to others. To the individual making investment decisions, it is important to understand what caused the performance. Were there extra benefits from market timing or only extra transaction costs? Was stock selection superior?

Portfolio evaluation has evolved dramatically over the last 40 years. The acceptance of modern portfolio theory has changed the evaluation process from crude return calculations to rather detailed explorations of risk and return and the sources of each. Furthermore, 40 years ago, evaluation was not an integral part of many organizations. This has changed (in part from external pressure) so that at this time, most investment organizations incorporate evaluation as an integral part of their decision-making process.

This chapter discusses portfolio performance evaluation starting with the simplest concepts of risk and return and tracing the evolution of the science through the most recent thinking about multiple sources of risk. Major empirical results, as well modern theory, are reviewed. We examine questions such as how well the industry has performed and managers who outperform index funds can be identified.

EVALUATION TECHNIQUES

The evaluation of portfolio performance is essentially concerned with comparing the return earned on some portfolio with the return earned on one or more other portfolios. It is important that the portfolios chosen for comparison are truly comparable. This means that they not only must have similar risk but also must be bound by similar constraints. For example, an institution that restricts its managers to investing in bonds rated AA or better should not evaluate its managers by comparing their performance to the performance of portfolios that are unconstrained. Although such a comparison would be useful in evaluating the relevance of the constraint, it would not be relevant for evaluating the manager.

Often the return earned by a fund is compared to the return earned by a portfolio of similar risk. In other comparisons an explicit risk–return trade-off is developed so that comparisons can be made across funds with very different risk levels. In either case, it is necessary to be more precise about what is meant by risk and return.

Measures of Return

In earlier chapters, when we computed return, we calculated the capital gains plus dividends from an initial investment. Thus, if a security paid dividends of $3.00 and had a capital gain of $7.00 on an investment of $100, the return was

$$\frac{7+3}{100} = 0.10 = 10\%$$

The 10% return was the return over the period in which the capital gain occurred.

When evaluating a portfolio, generalizing our simple idea of return requires care. A problem occurs because there are many inflows and outflows of funds to the portfolio, and very different amounts of money are invested at different points in time. To illustrate this, consider the example shown in Table 26.1. The portfolio has increased in value by 10% in each period, yet the ending value is less than the beginning value because of net outflows. To determine the rate of return by comparing the ending value to the beginning value would not reflect these changes.

As a second example, consider Table 26.2. This table shows two different patterns of inflows and outflows. In both cases, over the entire period, the inflows equal the outflows. Furthermore, the rate of return earned by each fund is identical in each period. However, the ending value is very different because the fund manager of fund *A* had the good luck to have the funds in the period that was highly profitable.

If we just looked at the ending value compared to the beginning value over the full period, fund *A*'s performance would look superior. However, the period-by-period return is identical and (ignoring risk for the moment) so is the manager's performance. *Unless* the inflows and outflows are under the control of the manager (and in most cases they are not),

Table 26.1 Hypothetical Inflows and Outflows

	Period			
	0	1	2	3
1. Value before inflow or outflow	$100	$110	$231	$55
2. Inflow (outflow)	0	$100	($181)	
3. Amount invested	$100	$210	$ 50	
4. Ending value	$110	$231	$ 55	

Table 26.2 Cash Flows and Returns for Two Funds

	Period			
	0	1	2	3
Rate of return earned by each manager	20%	−10%	10%	
Fund *A*				
1. Value before inflow or outflow	100	240	126	$138.60
2. Inflow (outflow)	100	(100)	0	0
3. Amount invested	200	140	126	
4. Ending value	240	126	138.60	
Fund *B*				
1. Value before inflow or outflow	100	120	198	$107.80
2. Inflow (outflow)	0	100	(100)	0
3. Amount invested	100	220	98	
4. Ending value	120	198	107.80	

the manager should not be rewarded or penalized for the good or bad fortune of having extra funds available at a particular time.

We eliminate the effect of having different amounts of funds available if we calculate the rate of return in each time period and then compound the return to determine it in the overall period. When the rate of return is calculated this way, it is called the *time-weighted rate of return.* For fund *A* the return in the first period is (240 − 200)/200 = 20%. In the second period the return is (126 − 140)/140 = −10%. In the third period the return is (138.60 − 126)/126 = 10%. The overall return is the product of 1 plus each of the three one-period returns minus 1, or (1.20)(0.90)(1.10) − 1 = 0.188, or 18.8%. This return is the same for *A* and *B.* Because the manager's performance was identical, this is appropriate. Also, in the first example, the time-weighted rate of return would show the actual 10% return that was earned for each period. It would not penalize the manager for the net outflows encountered.

To calculate the time-weighted rate of return requires knowledge of the value of the fund anytime there is a cash inflow or an outflow. For a fund with frequent transactions, this involves substantial calculations. If the inflows and outflows are not related to the market performance, then less frequent calculations may yield a reasonable approximation. Often funds are sold in units. Inflows and outflows affect the number of units, but any one unit reflects the same initial investment. In this case, tracing the performance of one unit is equivalent to determining the time-weighted rate of return. Having examined return, it is necessary to look at risk.

Measures of Risk

There are two possible measures of risk that can be used: total risk or nondiversifiable risk. Consider a college endowment fund. Clearly the appropriate risk is the risk on the total assets. The college will find very little comfort in the fact that part of the risk could be diversified away if it held other assets when the portfolio under consideration contains its total assets. As an alternative, consider the pension fund of a large corporation. For example, at one time AT&T allocated its pension funds to 125 separate managers. The contribution to the risk of the pension fund as a whole from the portfolio under supervision of any of these managers is primarily the nondiversifiable risk. AT&T, in evaluating its managers, should look at return relative to nondiversifiable risk.

As discussed in earlier chapters, total risk is normally measured by standard deviation of return, whereas nondiversifiable risk is normally measured by the beta coefficient.

Table 26.3 Comparison of Investment Performance of Mutual Funds and Random Portfolios (Jan. 1960–June 1968)

	Number in Sample		Mean Beta Coefficient		Mean Return	
Risk Class	Mutual Funds	Equally Weighted Random Portfolios[a]	Mutual Funds	Equally Weighted Random Portfolios	Mutual Funds	Equally Weighted Random Portfolios
Low risk ($\beta = 0.5-0.7$)	28	17	0.614	0.642	0.091	0.128
Medium risk ($\beta = 0.7-0.9$)	53	59	0.786	0.800	0.106	0.131
High risk ($\beta = 0.9-1.1$)	22	60	0.992	0.992	0.135	0.137

[a]Approximately the same number in a group for each of the variants.
Source: Friend, Blume, and Crockett (1970).

Having discussed risk and return, it is appropriate to look at techniques for examining portfolio performance.

Direct Comparisons

As discussed before, one way to compare portfolios is to examine the return earned by alternative portfolios of the same risk. This is the procedure used by Friend, Blume, and Crockett (1970) in their examination of mutual funds. Mutual funds have been evaluated by academics more than any other group of investment vehicles. This attention, which may well be unwelcome, is due primarily to the fact that data on mutual funds' portfolios are publicly available. Throughout this chapter we illustrate the discussion of performance measurement with reference to mutual fund studies.

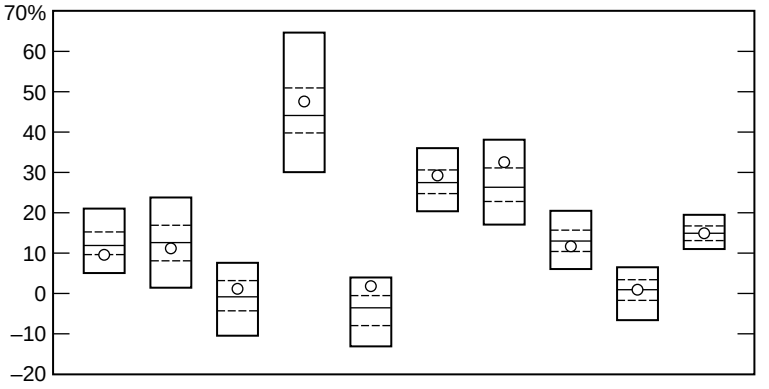
Table 26.3 from the Friend, Blume, and Crockett study shows the mean return earned by a group of mutual funds compared to randomly generated portfolios. In this table, beta was used as a measure of risk. The mutual funds were divided into three risk categories (high, medium, and low risk). Random portfolios with risks approximating the risk of the mutual funds were generated. The columns under mean beta show how closely they matched. The last two columns show the return on each group of random portfolios and mutual funds. In this period and for this measure, mutual funds did worse than randomly selected portfolios.

Friend, Blume, and Crockett repeat the analysis using variance as a measure of risk. Once again, mutual funds underperform randomly selected portfolios of the same risk. They also show that the population of securities, or more specifically, the weighting of the securities in random portfolios, can affect the evaluation results. Other characteristics, besides beta and standard deviation, can affect the valuation results. This will become clear when we examine multi-index and multiattribute measures of performance later in this chapter.

Most professional evaluation services chose as a benchmark not random portfolios but rather the performance of portfolios administered by other managers. Table 26.4a and 26.4b are part of the report of one of the large services that evaluates fund managers (usually pension funds).[1] Table 26.4a shows the return earned by the manager over the last year compared with the return of other fund managers. The circle shown in the chart represents the

[1]The major fund evaluation services such as S.E.I., Wilshire, and Barra have similar reports. We selected one at random for illustrative purposes.

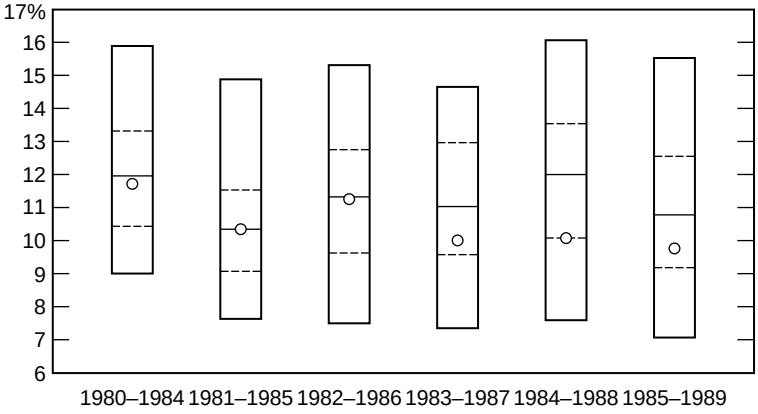
Table 26.4a Return



	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
5th Percentile	20.6	23.4	7.1	64.5	3.4	35.7	38.0	20.2	6.0	18.8
25th Percentile	14.7	16.3	2.7	50.8	–0.7	30.1	30.9	15.3	3.0	16.0
Median	11.4	12.1	–1.0	44.1	–3.7	27.3	26.1	12.5	0.6	14.5
75th Percentile	9.0	7.6	–4.4	39.5	–8.0	24.5	22.5	10.3	–1.7	12.8
95th Percentile	5.0	1.1	–10.6	29.8	–13.3	20.1	16.6	5.9	–6.9	10.8
Fund H1874	9.2	11.2	0.3	42.2	1.5	29.5	32.2	11.5	0.9	14.1
Percent rank	73	58	41	37	12	30	18	60	44	55

return of the manager in each year relative to other managers. The solid line in the middle of each rectangle represents the return of the median (50th percentile) manager. The upper and lower solid lines forming the rectangle represent the return of the 5th percentile and 95th percentile, respectively. Thus 90% of the managers lie within the rectangle. Finally, the

Table 26.4b Risk (Standard Deviation)



	1980–1984	1981–1985	1982–1986	1983–1987	1984–1988	1985–1989
5th Percentile	15.9	14.9	15.5	14.8	16.3	15.7
25th Percentile	13.5	11.6	12.9	13.1	13.7	12.6
Median	12.0	10.3	11.4	11.1	12.1	10.8
75th Percentile	10.4	9.1	9.5	9.6	10.1	9.2
95th Percentile	9.0	7.5	7.5	7.4	7.6	7.0
Fund H1874	11.7	10.4	11.3	10.0	10.1	9.7
Percent rank	53	49	52	70	75	67

dashed lines represent the return for the 75th and 25th percentile, respectively. This same information is presented numerically at the bottom of the chart. This type of information shows how well all managers did (by the position of the rectangle) and how well the manager being evaluated performed relative to other managers (by the position of the circle).

Table 26.4b shows similar information about the fund's risk. Table 26.4b compares the fund's total risk as measured by the standard deviation of return. Once again the circle represents the fund's performance and the rectangle encompasses 90% of all funds' standard deviations. Generally, both standard deviation and systematic risk as measured by beta are used as measures of risk. The same service will present comparisons using each measure separately.

Note that unlike the analysis of Friend, Blume, and Crockett, the return comparisons in Table 26.4a are not generally being made between funds of the same risk. Thus, although both return and risk measures are included as part of all evaluation services, it is often difficult to form an overall opinion about fund performance. Only in the two cases where risk and return are both adverse or good is it possible to form an overall opinion.

For example, if the performance evaluation indicated that the fund had a high risk relative to other funds and the return was consistently below average, the fund would be considered undesirable. Similarly, if the risk was consistently below average and the return consistently above average, the fund would be considered very desirable. Usually, however, there is no consistent pattern of return over time, and often the risk pattern varies as well. Thus the return pattern cannot be used to form an overall opinion about fund performance. The risk information may be useful in determining whether the manager has followed guidelines on risk. For example, if the manager was instructed to follow a strategy with a lower standard deviation than the average fund, was this policy in fact followed?

We have just shown that performance can be measured by comparing the returns of any portfolio with the return on other portfolios while examining risk. Performance measures developed since this early work combined attributes of risk and return into a single number. These measures can be divided into those that employ a single source of risk and those that employ multiple sources of risk. All make assumptions about capital markets that were not needed for the preceding analysis.

One-Parameter Performance Measures

Three different one-parameter performance measures have been proposed in the literature and are widely used in practice. We discuss each measure in turn. These measures differ in their definition of risk and their treatment of the ability of the investor to adjust the risk level of any fund in which she might invest. All of these measures implicitly make the assumption that the investor can both lend and borrow at the risk-free rate of interst.

The Excess Return to Variability Measure Consider the original portfolio problem. Figure 26.1 plots the return risk opportunities with riskless lending and borrowing.

As shown in Chapter 5, all combinations of a riskless asset and a risky portfolio lie along a straight line (in expected return standard deviation space) connecting the riskless asset and the risky portfolio. Thus the line $R_F A$ represents mixtures of the riskless asset and risky portfolio A, and $R_F B$ represents mixtures of the riskless asset and risky portfolio B. As we argued earlier, all investors would prefer portfolio A to B because combinations along $R_F A$ always give a higher return for the same risk. This idea can be and has been used for mutual fund evaluation.

Consider Figure 26.2. Portfolio A is being compared to portfolio B. If a riskless rate exists, then all investors would prefer A to B because combinations of A and the riskless asset give higher returns for the same level of risk than combinations of the riskless asset

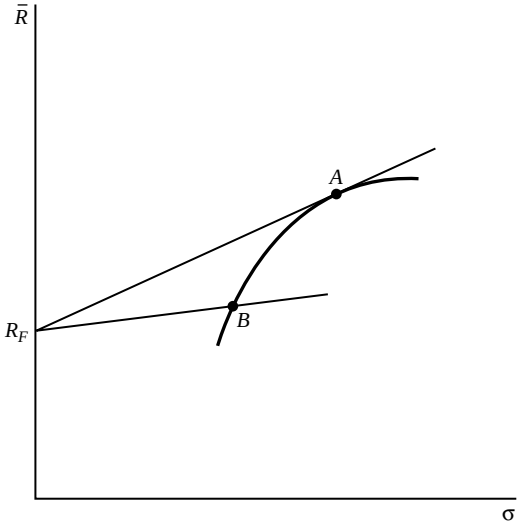


Figure 26.1 Combinations of a riskless asset and a risky portfolio.

and *B*. All combinations of any portfolio and the riskless asset lie in a ray that intersects the vertical axis at R_F. The preferred portfolio is that which lies on the ray passing through R_F, which lies furthest in the counterclockwise direction. In Figure 26.2 the portfolios are ranked alphabetically. Stating that the preferred portfolio lies on the most counterclockwise ray is equivalent to stating that the slope of the ray is the highest. In Chapter 6 we showed that the slope of the line was $(\bar{R}_p - R_F)/\sigma_p$. This ratio is one of the measures first utilized in portfolio evaluation and is called the Sharpe measure. An examination of the ratio shows that funds are ranked by the fund's return above the risk-free rate (excess return) divided by the standard deviation of return. This ratio is often referred to as an excess return to variability measure.

Figure 26.3 is a plot of individual mutual fund performance and the Dow-Jones Industrial index as presented in Sharpe's classic 1966 article. The ray connecting the risk free asset and

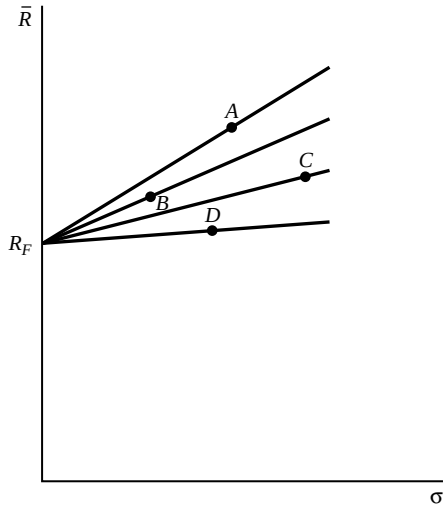


Figure 26.2 Combinations of a riskless asset and some mutual funds.

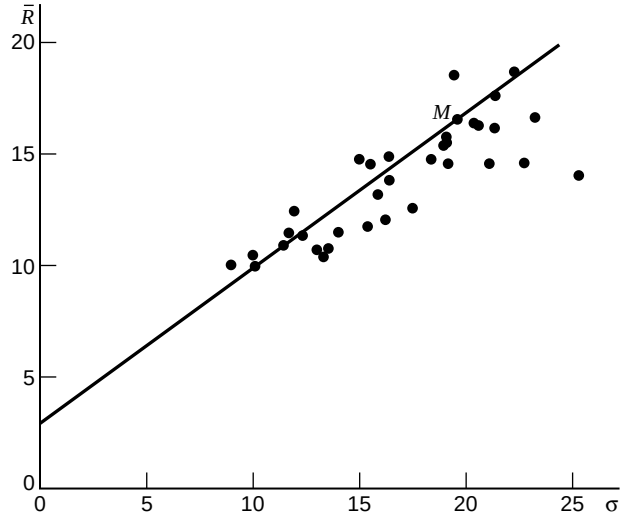


Figure 26.3 Funds in expected return standard deviation space.

the Dow-Jones index is shown in the diagram. Most of the mutual funds have a lower reward to variability index than the Dow-Jones index. This implies that most mutual fund managers in this period did worse than they would have done if they had simply invested in the Dow-Jones index and lent or borrowed to obtain their preferred risk.

The Sharpe measure looks at the decision from the point of view of an investor choosing a mutual fund to represent the majority of his investment. An investor choosing a mutual fund to represent a large part of her wealth would likely be concerned with the full risk of the fund, and standard deviation is a measure of that risk. Furthermore, if the investor desired a risk different from that offered by the fund, he would modify the risk by lending and/or borrowing. The relevant definition of performance may change if the problem is examined from the point of view of a fund manager or an investor evaluating the performance of a part of the total portfolio. This leads directly to our second measure of performance.

The sharpe ratio has a counterpart when nondiversifiable risk (beta) is chosen as the measure of risk. This may be more appropriate if one manager among many is being evaluated.

Consider portfolios in expected return beta space. It is easy to show that all combinations of a riskless asset and a risky portfolio lie on a straight line connecting them. Furthermore, the slope of the line connecting the risky asset A and the risk-free rate is $(\bar{R}_A - R_F)/\beta_A$. Once again, an investor would prefer the portfolio on the most counterclockwise ray emanating from the riskless asset.[2] In Figure 26.4 the portfolio ranking is alphabetical.

[2]Designate the beta on a portfolio of the riskless asset and portfolio A as β_p. Designate the beta on portfolio A as β_A and the beta on the riskless asset as β_F. The beta on a portfolio is a weighted average of the beta on the individual securities. Thus $\beta_p = X\beta_A + (1 - X)\beta_F$. But the beta on a riskless asset is zero or $\beta_F = 0$. Therefore $X = \beta_p/\beta_A$. The expected return on a portfolio is a weighted average of the expected return on the individual assets. Thus $\bar{R}_p = X\bar{R}_A + (1 - X)R_F$. Substituting β_p/β_A for X, we have

$$\bar{R}_p = \frac{\beta_p}{\beta_A}\bar{R}_A + \left(1 - \frac{\beta_p}{\beta_A}\right)R_F$$

Rearranging yields

$$\bar{R}_p = R_F + \left(\frac{\bar{R}_A - R_F}{\beta_A}\right)\beta_p$$

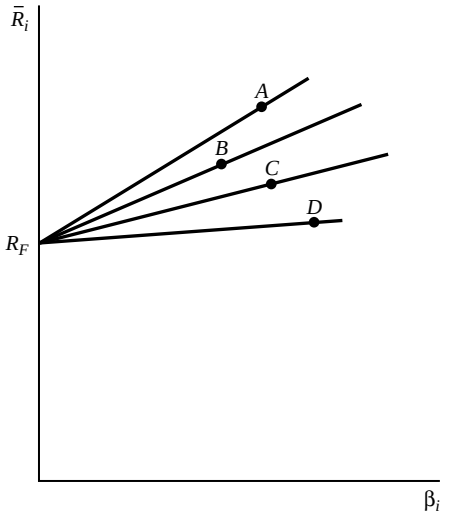


Figure 26.4 Treynor measure.

The Treynor Measure: Excess Return to Nondiversifiable Risk This measure of portfolio performance was first suggested by Treynor (1965) and is often called the *Treynor measure.* The final risk measure examines differential return when beta is the risk measure.

The Jensen Measure: Differential Return When Risk Is Measured by Beta Consider the line connecting the riskless rate and the market portfolio. A manager could obtain any point along this line by investing in the market portfolio and mixing this with the riskless asset to obtain the desired risk level. If the manager's choice is to actively manage the fund, then one measure of the manager's performance is the difference in return earned by actively managing the fund, compared to what would have been earned if the manager had passively invested in the market portfolio and riskless asset to achieve the same risk level. The slope of the line connecting the riskless asset and the market portfolio is $(\bar{R}_M - R_F)/\beta_M$, and the intercept must be the riskless rate. The beta on the market portfolio is 1. Thus the equation of the line is

$$\bar{R}_p = R_F + \left(\bar{R}_M - R_F\right)\beta_p \tag{26.1}$$

The differential return is the actual return less the return on the portfolio of identical beta, but lying on the line connecting the riskless asset and the market portfolio. This return is calculated, using the equation stated previously, along with the beta of the portfolio being evaluated.

Assume the market return is 10% and the risk-free rate is 5% and the beta on the portfolio being evaluated is 0.8. Then a mixture of the market portfolio and the riskless asset to obtain a beta of 0.8 would have an expected return of

$$\bar{R}_p = 5 + (10 - 5)(0.8) = 9\%$$

The differential return is the difference between the return on the portfolio and the 9% just calculated.

Table 26.5 Performance Summary—All Funds with Complete Data for 1960–1969 Period

			Average Values (Unweighted)				
Evaluation Period	Beta Range	No. Funds	Number of Observations (months)	Monthly Fund Return (%/month)	Average Beta	Monthly Market Return (%/month)	Differential Return
Jan. 1960	0−0.4	3	120	0.43	0.23	0.77	0.007
to	0.4−0.8	35	120	0.63	0.68	0.77	0.004
Dec. 1969	0.8−1.0	44	120	0.79	0.91	0.77	0.066
	1.0−1.2	30	120	0.86	1.07	0.77	0.056
	1.2+	13	120	1.05	1.33	0.77	0.130
	Total	125	120	0.78	0.91	0.77	0.051

This measure was first proposed by Jensen (1968) and is often referred to as the Jensen performance index. As an illustration of its use, consider Table 26.5 taken from the Security and Exchange Commission study of mutual funds. As seen from this table, mutual funds over the period 1960–1969 seemed to outperform the passive strategy.

The Jensen measure has a special appeal because of its relationship to the capital asset pricing models (CAPMs) discussed in Chapters 13–15. We presented the Jensen model as a comparison between the return on the mutual fund and the return on a portfolio constructed by mixing the riskless asset and the market portfolio to obtain the same risk. There is an alternative way of viewing the Jensen measure. Equation (26.1) is, of course, the capital asset pricing line discussed in Chapter 13. The differential return can be viewed as the difference in return earned by the fund compared to the return that the capital asset pricing line implies should be earned. Viewed in this way, the Jensen measure becomes a special case of a large number of measures that could be used.

In practice most users estimate Jensen's measure by running the time series regression

$$R_p = \alpha_p + R_F + \beta_p(R_m - R_F) + e_p$$

where alpha is the estimate of Jensen's differential return. In fact, in common parlance, alpha has been used as a term to describe differential performance, using any one of a host of performance models.

A MANIPULATION-PROOF PERFORMANCE MEASURE

As explained in Chapter 4, the use of standard deviation as a measure of risk is justified where return distributions are symmetric. In that case, ranking alternatives by standard deviation gives the same result as ranking by alternative downside risk measures that are perhaps intuitively more appealing. However, active trading, or the introduction of derivative securities into the portfolio, can introduce important asymmetry into the return distribution. A good example is where a trader supplements an equity portfolio with out-of-the-money calls and puts written on the same assets. In this case, the resulting negative skew implies that standard deviation is an inadequate measure of the downside risk the trader is assuming. Indeed, if the options are deep enough out of the money, the reduction in upside potential through the written calls can lead to a reduction in standard deviation. At the same time the option premia supplement portfolio returns. As a result the Sharpe ratio rises even

though this trading strategy involves no special knowledge or information unavailable to the rest of the market. Goetzmann, Ingersoll, Spiegel, and Welch (2007) identify such behavior as "information-less" trading. While it is easy to see that the Sharpe ratio can be artificially augmented in this way,[3] practically all portfolio performance measures can be manipulated through information-less trading that is not based on superior information or investment skill. Goetzmann, Ingersoll, Spiegel, and Welch (2007) identify even more sophisticated information-less trading strategies based on changing portfolio weights through time. As a solution to the performance manipulation problem, they propose a manipulation-proof measure, which cannot be gamed by options or dynamic trading strategies of this nature.

A manipulation-proof measure must have several characteristics: first, it must rank portfolios based on investor preferences; second, it cannot reward information-less trading; third, it should work for small and large portfolios alike; and fourth, it should be consistent with standard market equilibrium models. All of these conditions are met by a measure based on an average power utility model:

$$\hat{\Theta} \equiv \frac{1}{(1-\rho)\Delta t} \ell n\left(\frac{1}{T}\sum_{t=1}^{T}\left[(1+r_t)/(1+r_{ft})\right]^{1-\rho}\right)$$

where the $\hat{\Theta}$ statistic is an estimate of the portfolio's premium return after adjusting for risk. That is, the portfolio has the same score as does a risk-free asset whose continuously compounded return exceeds the interest rate by $\hat{\Theta}$. Here T is the total number of observations, and Δt is the length of time between observations. These two variables serve to annualize the measure. The portfolio's (unannualized) rate of return at time t is r_t, and the risk-free rate is r_o. The coefficient ρ should be selected to make holding the benchmark optimal for an uninformed manager. The measure is easy to calculate. As an example, consider a fund with monthly returns of -10%, 5%, 17%, and -3% when the continuously compounded monthly risk-free rate is 1%. If ρ is 2, then $\hat{\Theta} = 6.6\%$, and the fund has the same score as a risk-free asset with an annual rate of return of 18.6%. For ρ equal to 3, then $\hat{\Theta} = 1.1\%$, and the fund is equivalent to a risk-free asset returning 13.1%. The score is higher when $\rho = 2$ because its risk is not so heavily penalized.

TIMING

Another aspect of mutual fund performance that has been studied is timing. The question is how successful managers have been in timing the market, and also how timing is measured. Timing involves a change in the sensitivity of a portfolio to one or more systematic influences in anticipation of future movement in these influences. In the single index model we are now examining, this involves changing the beta on the portfolio in anticipation of future movements in the market index. There are three ways a manager can change the beta in his portfolio. The manager can sell stocks and buy debt instruments if he believes that the stock market will perform poorly. Second, the manager can sell high beta stocks and buy low beta stocks, if he believes the market will underperform. A third way, and one which involves less transaction costs, is for the manager to write stock index futures. Any of these three ways changes the beta on the portfolio.

The easiest way to examine the effectiveness of attempts to market time is to graphically examine market movements versus the bond–stock mix or average beta. Figure 26.5 is an

[3]In private communication, Jonathan Ingersoll shows that in a complete market environment, any portfolio strategy that generates payoffs that are uniformly concave (convex) relative to a buy-and-hold strategy will generate a Sharpe ratio larger (smaller) than that of a buy-and-hold strategy. Among other things, this implies that rebalancing to a constant long-term asset mix portfolio will lead to elevated Sharpe ratios, while protective put or portfolio insurance strategies will serve to diminish the Sharpe ratio below that of a buy-and-hold strategy.

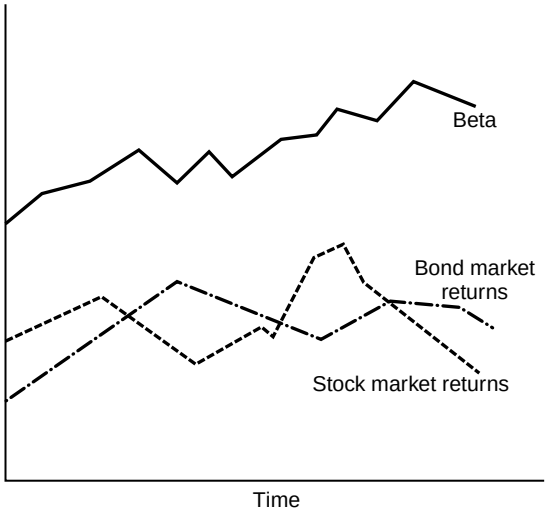


Figure 26.5 Beta and security returns.

example using betas. For this fund, very little evidence of successful timing is present. If the fund has a well-specified policy regarding the average beta or the bond–stock mix, then it is more illuminating to examine the relationship between deviations from the policy and changes in the market. Figure 26.6 is an example of this type of analysis.

Another measure of a manager's timing ability is to look at a plot of portfolio beta or bond–stock mix compared to the market return. If there is significant timing ability, then there should be a relationship between these variables, and this should be apparent from the plot.

A third way to measure market timing is to look directly at the fund return compared to the market return. If the fund did not engage in market timing, then the average beta on the overall portfolio should be fairly constant. If there was no diversifiable risk in the portfolio, then the portfolio return would be a constant fraction of the market return. A plot of market return

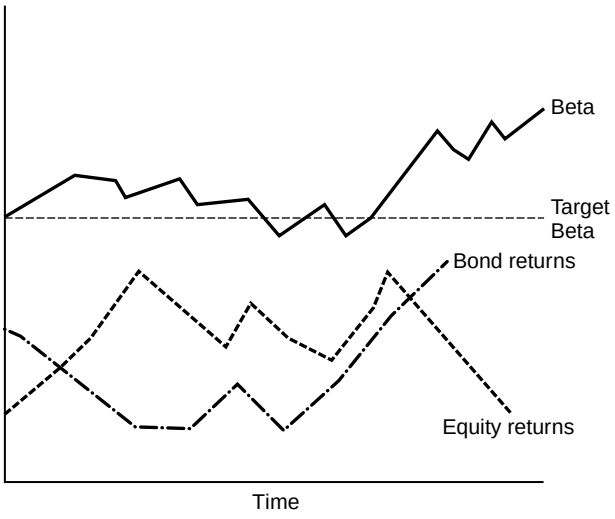


Figure 26.6 Measuring timing.

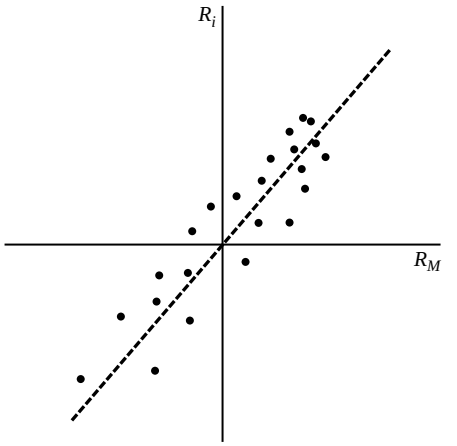


Figure 26.7 Returns for manager without timing.

compared to portfolio return would be a straight line. On an actual portfolio, there is usually some diversifiable risk and some changes in both beta and the bond–stock mix. If there was no successful timing, then these differences would simply cause the relationship between market return and portfolio return to be a scatter of points around a straight line, such as that shown in Figure 26.7. Assume a fund was able to engage in successful timing through changing beta. In this case, when the market increased substantially, the fund would have a higher than normal beta and would tend to do better than it would have otherwise done. This would cause the points to be above the normal line in Figure 26.7 for large market changes. Likewise, if the manager were able to anticipate a market decline, he or she would reduce the beta and have a portfolio that declined less than it would otherwise. This would mean that for low market returns, points would tend to scatter above the normal relationship. The points above the normal relationship for low and high market returns would give a curvature to the scatter of points if there were successful timing. An example is shown in Figure 26.8. Treynor and Mazuy (1966) utilized this to analyze the timing ability of mutual funds. They found that only one fund out of the 37 they examined exhibited any significant timing ability. They did not examine the odds that one would observe one such fund when no timing ability existed.

The Treynor and Mazuy procedure to test for curvature is to fit a quadratic curve to the performance data. The following multiple regression is run:

$$\left(R_{it} - R_{Ft}\right) = a_i + b_i\left(R_{mt} - R_{Ft}\right) + c_i\left(R_{mt} - R_{Ft}\right)^2 + e_{it}$$

where

R_{it}	is the return on fund i in period t
R_{mt}	is the return on the market index in period t
R_{Ft}	is the riskless asset
e_{it}	is the residual return of fund i in period t
a_i, b_i, and c_i	are constants

If the relationship between the fund's returns and the market's returns are as shown in Figure 26.7, then a straight line will best fit the scatter of points. In this case the addition of a squared term will not improve the fit and c_i will be zero. If the relationship between the fund and the market is as shown in Figure 26.8, then the addition of a squared term

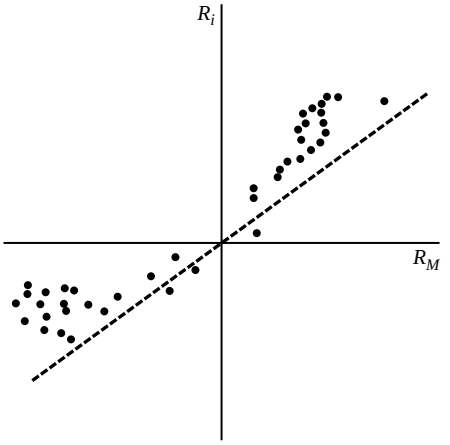


Figure 26.8 Returns for manager with timing.

(which results in a curved shape) will improve the fit, and c_i will be positive. Thus c_i is a measure of the fund's timing ability.

An alternative way to analyze market timing is to fit two separate lines. One line is fit for the observations when the market outperforms the riskless asset (up markets) and the other line is fit when the market underperforms the riskless asset (down markets). A manager with market timing should have a high up-market beta and a low down-market beta. This method is presented in Henriksson and Merton (1981).

This idea can be implemented by estimating the parameters in the following regression:

$$\left(R_{it} - R_{Ft}\right) = a_i + b_i\left(R_{mt} - R_{Ft}\right) - c_i D\left(R_{mt} - R_{Ft}\right) + e_{it}$$

where

$D = 0$ if $R_{mt} - R_{Ft} \geq 0$ an up market

$= 1$ if $R_{mt} - R_{Ft} < 0$ a down market

To illustrate why this works, consider what the equation looks like for different values of $R_{mt} - R_{Ft}$:

$R_{mt} - R_{Ft}$	EQUATION
+	$R_{it} - R_{Ft} = a_i + b_i(R_{mt} - R_{Ft}) + e_i$
0	$R_{it} - R_{Ft} = a_i$
−	$R_{it} - R_{Ft} = a_i + (b_i - c_i)(R_{mt} - R_{Ft}) + e_i$

Examining these equations shows that b_i is the up-market beta and $(b_i - c_i)$ is the down-market beta; c_i is the difference between the up-market beta and the down-market beta. A successful market timing will have a positive c_i. If c_i is statistically significant, it is some indication that the result is not due to luck but rather to skill.

These standard timing tests that typically rely on monthly data can be improved. Managers may make their timing decisions at higher frequencies. Goetzmann, Ingersoll, and Ivkovich (2000) shows that the mismatch between data frequency and manager activity can dramatically weaken standard return–based timing methods when monthly data are used. They propose a correction that leads to more powerful tests and may be employed with a single-factor or multifactor model.

One problem with using beta to measure performance becomes particularly acute when measuring timing. Active management involving timing in particular means that management action causes beta to be changed over time. Thus the estimate of beta using time series is measured over different beta regimes as constructed by the portfolio manager at different points in time. This has led to the development of new measures of timing.

HOLDING MEASURES OF TIMING

Elton, Gruber, and Blake (2011b), Daniel, Grinblatt, Titman, and Wermers (1997), and Jiang, Yao, and Yu (2007) use holdings data to estimate mutual fund betas and to measure timing. Since the betas on a portfolio are a weighted average of the betas on the securities that comprise the portfolio, there is an alternative time series way to estimate a mutual fund's beta. They can be estimated by first estimating each security's betas, then using holdings data to obtain security proportions, and finally using the product of security betas and proportions to get the mutual fund betas. The advantage of this approach is that it avoids the following problem: if management is changing the composition of a portfolio over time (e.g., because it is engaging in timing), the betas on the fund from a time series regression of fund returns will be poorly specified. Using holdings data at each point in time that holdings are observed provides a direct estimate of beta.[4]

Elton, Gruber, and Blake (2011b) measure timing using a method parallel to how alpha is measured. They measure timing as the difference in performance between the actual beta and the target beta (specified below) at the end of the period times the return in the next period. In equation form for any index,

$$\text{Timing} = \sum_{t=1}^{T} \frac{(\beta^* - \beta_t)R_{pt}}{T}$$

where

1. β_{At} is the actual beta in period t
2. β^*_t is the target beta in period t
3. T is the number of time periods
4. Other terms standard

This measure captures whether the fund deviated from the target beta in the same direction as the return on the index deviated from its normal pattern. Does the fund increase its beta when index returns are high and decrease when index returns are low?

There are several possibilities about what to use for a target beta. For a plan sponsor trying to evaluate a fund that professes to be a timer and has an agreed-upon normal beta, the target beta might be the agreed-upon beta. For an outside observer, the average beta over time might be a reasonable choice. Finally, if one believes the market can be forecasted and the forecasting procedure is widely known, and if one also believes that the manger should not get credit for using this public information, then the target beta could be the forecasted beta. For example, if one believes that the market can be forecasted by the dividend price ratio and that the manager should not be given credit for changing beta in response to changes in the dividend price ratio, then beta forecasted by the dividend price ratio could be used as a target beta. Ferson and Schadt (1996) discuss how to capture

[4]The methodology here has been applied to each index in the multi-index performance measured described shortly.

changing beta from public information when timing is measured using historical returns. The same idea can be used here.

The Elton, Gruber, and Blake (2011b) measure is similar in concept to one developed earlier by Daniel, Grinblatt, Titman, and Wermers (1997). As a target beta, Daniel, Grinblatt, Titman, and Werners use the beta from the previous period. The difference in beta is then the change in beta from the prior period. Finally, Jian, Yao, and Yu (2007) measure timing in a different manner. They develop a version of the Treynor and Mazuy measure where beta is measured from holdings data.

In comparing timing measures we have a strong preference for measures based on holdings data. The use of holdings data allows the researcher to observe the pattern of changes in betas and sector or industry weights over time. It also captures much more complicated timing strategies than, for example, assuming the manager switches between two betas depending on her market forecast. We also prefer thinking of timing as deviations from a target beta because this is how plan sponsors view timing. However, there is merit in asking whether a manager gained or lost due to changing beta from the level held in the prior period.[5]

MULTI-INDEX MODELS AND PERFORMANCE MEASUREMENT

As we discuss, any single-index model can lead to erroneous conclusions about the performance of any fund and the industry in general. This has led to a new set of metrics.

Multi-index Benchmarks Estimated Using Returns Data

Viewing a portfolio as a combination of the market and the riskless asset ignores other characteristics of the portfolio that affect performance. Merton (1973) suggests that an investor may be concerned with other influences such as inflation risk. Ross (1976) develops the arbitrage pricing model (APT), which shows how returns can depend on other systematic influences. These developments lead to researchers considering a generalization of Jensen's model:

$$R_{pt} - R_{Ft} = \alpha_p + \sum_{k-1}^{K} \beta_{pk} I_{kt} + e_{pt}$$

where the Is represent influences that systematically affect returns and the βs sensitivity to these influences.

What Is or systematic influences should be used in the model? The literature on performance measurement has employed several methods of determining the Is. They include

1. indexes based on a set of securities that are hypothesized as spanning the major types of securities held by the mutual funds being examined
2. indexes based on a set of portfolios that have been shown to explain individual security returns
3. indexes extracted from historical returns using forms of statistical analysis (factor analysis or principal components analysis)

These approaches are described subsequently.

[5]In a multifactor model, measuring timing requires measuring changes with respect to all factors that affect return, not just timing with respect to the market, as discussed in Elton, Gruber, and Blake (2011b). Managers changing sensitivity to factors other than the market will also affect market sensitivity, and it is the sum of all effects that determines the impact on returns from this decision. Likewise, managers, by changing market sensitivity, are also likely to be changing sensitivity to other factors, and again, it is the sum of all these effects that measures the impact of this decision on shareholder returns.

Indexes based on the major types of securities held by a fund. The first attempts to expand beyond the single-index model were performed by Sharpe (1992) and Elton, Gruber, Das, and Hlavka (1993). The motivation for EGD&H's development of a three-index model (the market, an index for small stocks, and an index for bonds) was the work of Ippolito (1989). Unlike earlier studies, he found that mutual funds had, on average, large positive alphas using Jensen's model. Furthermore, funds that had high fees tended to have higher alphas after fees. The period studied by Ippolito was a period when small stocks did extraordinarily well, and even after adjusting for risk, passive portfolios of small stocks had large positive alphas. Realizing that Ippolito's sample included many funds that invested primarily in mid-cap or small stocks and small-cap stock funds tend to have bigger fees explains Ippolito's results. By including indexes for small stocks and bonds (Ippolito's sample included balanced funds), the surprising results reported by Ippolito were reversed. Funds on average tended to have negative alpha, and those funds with high fees tended to perform worse than funds with low fees.

Simultaneously with the EGD&H exploring the return on plain vanilla U.S. stock funds, Sharpe (1992) was developing a multi-index model to explain the return on a much more diverse set of funds. He employed 16 indexes to capture the different types of securities that could be held by a wider set of funds.

The type of analysis performed by EGD&H and Sharpe not only produced better measurement of performance but allowed the user to infer, by observing the weights on each index, the type of securities held by the fund. This type of analysis has become known as return-based style analysis. It allows style to be inferred without access to individual fund holdings. EGD&H and Sharpe differ in the way they estimate their models. EGD&H use ordinary least squares, while Sharpe constrains each beta to be nonnegative and the sum of the betas to add to 1.[6] The advantage of Sharpe's approach is that the loading on each type of security can be thought of as a portfolio weight. The disadvantage is that by introducing additional constraints, the model does not fit the data as well.

Indexes based on influences that explain security characteristics.[7] While authors have continued to use security-based models, often adding indexes to better capture the types of securities held (e.g., foreign holdings), a particular form of multi-index model has gained wide acceptance. This model is based on Fama and French's (1996) findings that a parsimonious set of variables can account for a large amount of the return movement of securities. The variables introduced by Fama and French include, in addition to the CRSP equally weighted market index minus the riskless rate, the return on small stocks minus the return of large stocks, and the return of high book-to-market stocks minus the return of low book-to-market stocks (value–growth).

While the Fama–French model has remained a basic multi-index model used to measure portfolio performance, in many studies, two additional variables have sometimes been added. The most often-used additional index was introduced by Carhart (1997). Drawing on the evidence of Jegadeesh and Titman (1993) that stock returns, in part, can be predicted by momentum, Carhart added a new variable to the three Fama–French variables: momentum.

[6]Performance is estimated by Sharpe from a quadratic programming problem that minimizes the squared deviations from a regression surface given a set of linear constraints on the sign and the sum of betas.

[7]One and two may seem similar. The difference is that one incorporates the types of securities held by a fund, whereas two incorporates influences (which may be portfolios of securities) that are used because they explain security returns.

Momentum is usually defined as follows: the difference in return on an equally weighted portfolio of the 30% of stocks with the highest returns over the previous 12 months and a portfolio of the 30% of stocks with the lowest return over the previous 12 months.

The idea behind incorporating this index is a belief that past return predicts future return, and management should not be given credit for recognizing this. Later we will examine additional attempts to correct management performance for other types of publicly available information. Unlike indexes that represent sectors of the market such as large stocks, where index funds are readily available, the question remains as to whether management should be given credit for incorporating publicly available information into portfolio decisions. To the extent that vehicles do not exist to take advantage of this and the correct way to incorporate this information is not clear, a case can be made for not incorporating these indexes. In addition, Elton, Gruber, and Blake (2011a), when examining momentum using security data to measure portfolio betas, found tremendous instability in the year-to-year beta with respect to momentum, with some firms switching from following momentum to contrarian in successive years, suggesting many managers are not adopting the academic literature suggesting momentum predicts return.

Another addition to the Fama–French or Fama, French, and Carhart models is to add a bond index to the model. The index is usually constructed as the return on a long-term bond index minus the return on the riskless rate. Its introduction is intended to adjust for the fact that many managers hold long-term bonds in their portfolio and that these securities have characteristics not fully captured by the other variables in the Fama–French model. Failure to include this index means that funds which have bonds other than one month T-bills will have the difference in performance between the bonds they hold and T-bills reflected in alpha. The effect of this on performance has been documented in Elton, Gruber, and Blake (1996b).

In the last decade or so, many mutual funds that are labeled as U.S. equity funds have included foreign securities as part of their portfolio. To date, few researchers have included indexes to capture this. However, research using data from the last decade or so needs to be conscious of this and take it into account. In addition, the Fama–French growth factor may not adequately capture the impact of growth. In many studies the performance of growth funds is less well measured than the performance of other types of funds. Elton, Gruber, and Blake (1999) find that an index of the returns on growth mutual funds does a better job of explaining mutual fund performance than the growth measures used in most of the mutual fund literature.

Indexes extracted from historical returns. Another approach to identifying the appropriate indexes to use in the performance model is to use a form of statistical analysis (factor analysis or principal component analysis) to define a set of indexes (portfolios) such that the return on this set of portfolios best explains the covariance structure of returns and reproduces the past returns on securities and portfolios. Connor and Korajczyk (1986, 1988) present the methodology for extracting statistical factors from stock returns, and Lehman and Modest (1987) apply the statistical factors to evaluating mutual fund performance. This methodology continues to be used to evaluate mutual fund performance.

Performance Measurement Using Multi-index Models Most studies employing multi-index models and the Jensen measure use the α estimated from a multi-index model directly as a performance measure replacing the single-index alpha.

Sharpe has suggested an alternative to the traditional Sharpe measure called the generalized Sharpe measure that is an alternative to using alpha directly. In this measure a benchmark return replaced the riskless rate in the numerator of the traditional Sharpe

measure and is used to define the denominator. The benchmark is determined by estimating the betas from a time series regression of the returns on a fund versus a set of indexes spanning the relevant characteristics of the securities held by the fund. Define the benchmark as

$$R_{Bt} = \sum_{k=1}^{K} B_{pk} I_{kt}$$

Sharpe (1992) formulated the generalized Sharpe measure as the average alpha over the standard deviation of the residuals, or in equation form,[8]

$$\frac{\frac{1}{T}\sum_{t-1}^{T}\left(R_{pt} - R_{Bt}\right)}{\left[\frac{1}{T}\sum_{t=1}^{T}\left[R_{pt} - R_{Bt}\right]^2\right]^{1/2}}$$

Clearly, as Sharpe has pointed out, this is superior to the original Sharpe model for almost all purposes.

While the use of multi-index models estimated from a time series regression have been widely used to infer performance and style, several researchers have suggested using holdings data to correct potential weaknesses in time series estimation.

Using portfolio composition to estimate portfolio betas. The models discussed to this point estimate betas from a time series regression of portfolio returns on a set of indexes. One difficulty with this approach is that it assumes that betas are stable over the estimation period. However, if management is active, the betas on a portfolio may shift over time as management changes the composition of the portfolio. Because portfolio weighs changes as a function of management action, the estimates of portfolio betas from time series regression may not be well specified.[9] Potentially better measure of the betas on a portfolio at a moment of time can be estimated by combining the betas on individual securities with the weight of each security in the portfolio at that moment of time. This approach to estimating betas and alphas has been examined by Elton and Gruber writing with others (2010b, 2011a, 2011b) in three contexts: to forecast future performance, to discern timing ability, and to study management reaction to external phenomena. The results indicate significant improvement is obtained by estimating betas from portfolio holdings.[10] For example, they compared the forecasting ability of alpha when alpha was computed using betas calculated from security betas and holdings data (bottom-up) with alpha computed from running a time series regression using mutual fund returns (top-down). Bottom-up alphas better forecasted future performance whether future alphas were computed using bottom-up or top-down betas.

USING HOLDINGS DATA TO MEASURE PERFORMANCE DIRECTLY

A second approach to using holdings-based data was developed by Daniel, Grinblatt, Titman, and Wermers (1997). Daniel et al. formed 125 portfolios by first sorting all stocks into five groups based on market capitalization, then within each group, forming five

[8]Despite Sharpe's article describing and defending the generalized Sharpe ratio, industry practice and much of the literature of financial economics continues to use the original Sharpe ratio in evaluating performance. Note that Sharpe has the riskless rate as a variable in his benchmark. If one used the normal regression procedure, portfolio returns and index returns would need to be in excess return form.

[9]Wermers (2002) documents a significant amount of style drift for mutual funds over time.

[10]Two other studies have used this method of estimating betas in timing studies, but these will be reviewed later in this chapter under "Timing."

groups sorted by book-to-market ratios, and finally, within these 25 groups, five groups by momentum. Passive returns on each of the 125 portfolios are then calculated as an equally weighted average of the return on all stocks within each of the 125 groups. The benchmark return for any fund is found by taking each stock in a fund's portfolio and setting the benchmark return for each stock as the return on the matched cell out of the 125 cells described earlier. They then used the benchmark described earlier to measure security selection as follows:

$$\alpha_p = \sum_{i=1}^{N} w_{it}(R_{it} - R_{itB})$$

Here the weight w_{it} on each stock at the end of period is multiplied by the return on that stock in period t to $t + 1$ (R_{it}) minus the return that would be earned on a portfolio of stocks with the same book-to-market, size, and momentum (R_{itB}), and the result summed over all stocks in the portfolio. This approach, like the Fama–French Carhart approach, assumes we have identified the appropriate dimensions of return. It does not assume the linear relationship between characteristics and return inherent in a regression model. On the other hand, the cost of the approach in terms of data is great, and the comparisons are discrete in the sense that comparison is made to the average return in one of 125 cells rather than as a continuous variable.

Another approach to using portfolio composition to measure performance has become known as the weight-based measure of portfolio performance. The basis of this measure is the research of Cornell (1979) and Grinblatt and Titman (1989a, 1989b). Many portfolio holdings measures are based on comparing performance to what it would have been if the manager had not changed the weights. The idea is simple and appealing. If the manager increases weight on securities that do well in the future and decreases weights on securities that do poorly, he is adding value.

Almost all holdings-based metrics do not measure the performance an investor in the fund would achieve, but rather whether the manager adds value by her security selection. The exception to this is the holdings method employed by Elton, Gruber, and others (2010b, 2011a, 2011b), which can be used to measure performance both pre- and post expenses.

TIME-VARYING BETAS

The regression techniques described earlier assume that the sensitivities of a fund to the relevant characteristics remain constant over time. Using holdings data to estimate betas is one way of dealing with changing betas.

An alternative to using holdings data to estimate changing betas is to fit some functional form for how betas change over time.

CONDITIONAL MODELS OF PERFORMANCE MEASUREMENT, BAYESIAN ANALYSIS, AND STOCHASTIC DISCOUNT FACTORS

Two approaches have been set forth as a modification of the standard models of portfolio performance. The first recognizes that the risk sensitivity of any mutual fund can change over time due to publicly available information, while the second uses Bayesian techniques to introduce prior beliefs into the evaluation process. The philosophy behind conditional models of performance measurement is that sensitivity to indexes should change over time because return on these indexes is partially predictable. Furthermore, management should not be given credit for performance, which could be achieved by acting on publicly available

information that can be used to predict return. We already briefly discussed this philosophy when we examined the Carhart model.

In a broader sense, the extreme version of the conditional model says that superior performance occurs only if risk-adjusted returns are higher than they would be based on a strategy of changing sensitivity to indexes by using public information in a mathematically defined manner.

Ferson and Schadt (1996) develop one of the best-known and often-used techniques for conditional beta estimation. Their version of the traditional CAPM specifies that risk exposure changes in response to a set of lagged economic variables which have been shown in the literature to forecast returns. The model they specify is

$$R_{pt} - R_{Ft} = a_p + \beta_p(Z_i)(R_{Mt} - R_{Ft}) + e_{pt}$$

where $\beta_p(Z_t)$ is the value of the conditional beta (conditional on a set of lagged economic variables) at a point in time. These conditional betas can be defined as

$$\beta_p(Z_t) = \beta_{p0} + \beta_{p1}Z_t$$

where Z_t represents a set of conditioning variables. Ferson and Schadt use four lag variables as conditioning variables: T-bill rates, dividend yield, the slope of the term structure, and a measure of quality spread.[11]

Mamaysky, Spiegel, and Zhang (2007) take a different approach to measuring performance, with time-varying coefficients. Rather than hypothesizing a set of lagged variables that help to determine betas at a period in time, they used Kalman filters to determine the time pattern of betas and performance over time. This allows the pattern to be determined by a set of variables that are statistically estimated rather than hypothesized by the researchers.

BAYESIAN ANALYSIS[12]

A number of authors have used Bayesian analysis to continuously adjust the alpha resulting from a multi-index model. Baks, Metrick, and Wachter (2001) assume that an investor has prior beliefs concerning whether any manager has skill. They use this prior and the history of returns to compute the posterior α using Bayesian analysis.

Pastor and Stambaugh (2000) assume α multi-index model. First they divide their indexes into those that an investor believes are in a pricing model and those that are not (labeled nonbenchmark assets). Pastor and Stambaugh (2002) show that if nonbenchmark assets are priced by benchmark assets exactly, then αs are completely unchanged by the choice of an asset pricing model. However, if they are not priced exactly, different models will produce different estimates of alpha, and by incorporating a set of nonbenchmark passive portfolios on the right-hand side of the return regression, a better estimate of alpha is obtained. Pastor and Stambaugh assume investors have prior beliefs on how certain they are that they have correctly identified the correct asset pricing model and use Bayesian analysis to update these beliefs.[13]

[11]The generalization of the Fearson and Schott procedure to a multifactor model is straightforward. Christopherson et al. (1998) propose that α as well as the betas are conditional on a set of lagged variables.

[12]Stambaugh (1997) showed how movements of assets with long histories can add information about movements of assets with shorter histories; thus one reason to examine nonbenchmark assets is that they may have a longer history.

[13]The Pastor–Stambaugh framework was applied by Busse and Irwin (2006) to daily data.

STOCHASTIC DISCOUNT FACTORS

Several authors—Chen and Knez, 1996; Farnsworth, Ferson, Jackson, and Todd, 2000; Dahlquist and Soderlind, 1999—have tried to estimate stochastic discount factors and then have evaluated mutual funds as the difference between the funds' performance and the return on the fund if it earned the equilibrium return using the stochastic discount function. The idea is parallel to Jensen's alpha when the single-factor model is interpreted as the CAPM model.

WHAT'S A RESEARCHER TO DO?

In the prior sections we have discussed many models useful for measuring mutual fund performance. What advice can we give in choosing among them? First, except for evaluating index funds, single-index models are generally inappropriate for measuring mutual fund performance. There are long periods of time when small stocks or value stocks have outperformed the market. Most funds have exposure different from that of the market to these factors. Failure to account for this difference in exposure is likely to attribute differences in the performance of these factors to the manager's performance.

Second, measures have been developed which are used to examine the effect of changes in portfolio holdings as opposed to the performance of overall portfolios. We view the former as less important, both because it can only measure performance before fees and because it ignores the effect on performance of the securities that are not traded. All of the other techniques can be used to measure performance before fees or after fees. The former is of interest to see if management has stock selection ability. The second is the relevant measure for investors: does management add value for the investor? The key question is do specific mutual funds or mutual funds in general outperform properly designed passive portfolios?

Third, how should we estimate sensitivities: from a time series regression on the funds' returns (top-down) or from estimate of the securities returns and portfolio weights (bottom-up)? There are clearly large errors in individual security betas, but there is also significant reduction (canceling out of errors) when we move to the portfolio level and examine results over time. We find much better prediction of future performance using bottom-up versus using top-down estimates of beta whether future performance is judged using bottom-up or top-down models. Thus bottom-up betas are likely to be measuring real changes in mutual fund betas over time. Changing betas over time can seriously affect the estimates of betas from a time series regression on fund returns.

We cannot make a definitive statement about which multi-index model should be used for measuring whether mutual fund managers outperform passive portfolios, except that the model needs to capture the major factors that affect mutual fund performance. This means that a researcher needs to be conscious of the types of securities the funds in her sample hold and not simply rely on the overall classification, for example, stock funds. Many common stock funds hold bonds, and failure to correct for this means that the difference between the performance of long bonds and the riskless rate gets impounded in the alpha. Likewise, many funds in recent years have held international stocks, and this has to be recognized. How can a researcher choose among models? Grinblatt and Titman (1989b) had a very clever suggestion. They argue that index funds of any particular type should not show a positive alpha with respect to an appropriate model. With the large number of index funds, this is an easy way to check any multi-index model. In the appendix we show how results from a basic multi-index model can be combined with an APT model to measure and diagnose performance.

MEASURING THE PERFORMANCE OF ACTIVE BOND FUNDS

Although there has been a vast literature on models for evaluating stock mutual funds, the literature dealing with the performance of bond funds is much less developed. This is true despite the fact that bond funds constitute a significant proportion of mutual fund assets.

The first paper to present a detailed analysis of bond fund performance was Blake, Elton, and Gruber (1994). In this paper the authors employ regression models of the type discussed earlier, as well as the QPS version of this model developed by Sharpe (1992). Blake, Elton, and Gruber investigated a one-index model (either a general bond index or the submarket index that Morningstar identified as most like the bond fund), two three-index models, and a six-index model.

The six indexes were based on the major types of securities held by the fund and included an intermediate government bond index, a long-term government bond index, an intermediate corporate bond index, a long-term corporate bond index, a high-yield bond index, and a mortgage bond index. Unlike stocks, where performance seems extremely sensitive to the choice and definition of the indexes employed, the results for bond funds seem to be fairly robust across models, as long as three indexes are used. The three indexes needed were a general bond index, a high-yield index, and either a mortgage or term structure index.

A series of papers, particularly Elton, Gruber, and Blake (1995), Comer and Rodriguez (2006), and Chen, Ferson, and Peters (2010) continue the investigation of multi-index models to measure the performance of bond portfolios. While many of these papers used different indexes to measure performance, the general results indicate that the researcher should use at least three indexes: a general bond index, a risk index such as a high-yield index, and an index to measure option-like qualities such as a mortgage index. It is worth noting that performance results for bond funds are much more robust to the choices of competing indexes than they are for stock funds.

Having discussed alternative models to measure mutual fund performance, we now turn to the results from applying many of these models to both stock and bond funds.

THE PERFORMANCE OF ACTIVELY MANAGED MUTUAL FUNDS

Two aspects of mutual fund performance should be of key interest. How well have actively managed mutual funds done in general, and is there persistence in mutual fund performance? If the answer to the latter question is yes, then the question of interest becomes, can we identify actively managed mutual funds will outperform passively managed mutual funds?

HOW HAVE MUTUAL FUNDS DONE?

The evidence that actively managed mutual funds have on average negative alphas after fees and positive alphas before fees is very robust. The results of a representative set of studies are presented in Table 26.6. This table presents the average alphas found by a large set of researchers using very different ways of measuring alpha and measuring performance over different time periods.

The results of Panel A show that mutual funds after fees underperform passive portfolios by 65 to 200 basis points per year. In Panel B we see that actively managed mutual funds outperform passive portfolios before expenses. The results are quite interesting because they indicate that managers have selection ability but not enough to cover expenses. This

Table 26.6 Mutual Fund Performance Results (Annualized)

A. Articles Using Mutual Fund Returns (Post Expenses)		
		Average Performance
1.	Jensen (1968)	−1.1
2.	Lehman and Modest (1987)	Negative
3.	Elton, Gruber, Das, Hlavka (1993)	−1.59
4.	Gruber (1996)	−0.65
5.	Elton, Gruber, and Blake (1996b)	−0.91
6.	Forson and Schadt (1996)	+0.24
7.	Carhart (1997)	−1.98
8.	Pastor and Stambaugh (2002)	−0.86 to −1.25
9.	Elton, Gruber, and Blake (2003)	−0.91
10.	Fama and French (2010)	−0.83
11.	Elton, Gruber, and Blake (2011a)	Negative
B. Using Holdings Data (Before Expenses)		
1.	Grinblatt and Titman (1989a)	(slight positive)
2.	Grinblatt and Titman (1993)	2.00%
3.	Daniel, Grinblatt, Titman, and Wermers (1997)	0.77
4.	Wermess (2002)	0.71
C. Timing		
1.	Daniel, Grinblatt, Titman, and Wermers (1997)	Timing ability
2.	Busse (1999)	Timing ability
3.	Becker, Person, Myers, and Schill (1999)	No timing ability
4.	Bollen and Busse (2001)	Timing ability
5.	Kaplan and Sensoy (2005)	Timing ability
6.	Jiang, Yao, and Yu (2007)	Timing ability
7.	Elton, Gruber, and Blake (2011b)	No timing ability
8.	Ferson and Qian (2006)	No timing ability
D. Bond Funds		
l.	Blake, Elton, and Gruber (1994)	−0.51%
2.	Elton, Gruber, and Blake (1995)	−0.75% to −1.3%
3.	Corner and Rodriguez (2006)	−1.00 to −1.14%
4.	Chen, Ferson, and Peters (2010)	−0.70%

leads naturally to the next section of this chapter: can we identify those managers who have enough selection ability to more than offset the expenses they charge?[14]

Before turning to that subject we should mention that the results on [timing] are much more diverse (Table 26.6, Panel C). Of the major studies of timing ability, about half identify timing ability in the mutual fund industry and about half find no timing ability. For example, Jiang et al. (2007) find timing ability, while Elton, Gruber, and Blake (2011b) find no timing ability.

[14]The results for bond funds are similar. See, for example, Conner and Rodriguez (2000) and Chen et al. (2010).

THE PERSISTENCE OF PERFORMANCE

If mutual funds in general outperform index funds before expenses, this indicates that management has the ability to add value but that managers charge investors more than the value they add. This suggests two questions of interest: is there persistence in mutual fund performance, and can mutual funds be identified that will have positive alpha?

PERSISTENCE

A large number of studies have found persistence in mutual fund performance. Some of the key research is summarized in Table 26.7. Studies using a single-index model of performance sometimes find persistence. However, the results may be due to persistence in styles or sectors of the market (e.g., small stocks) not accounted for by the model. The studies cited in Table 26.7 that use different methodologies and examine different time periods all found that past performance has some predictive power for future performance.

Table 26.8 shows an example of persistence reported in Elton, Gruber, and Blake (2012).[15] This table shows that past fund performance is correlated with future fund performance results consistent with other studies noted earlier. Every study shows that poor performance predicts future poor performance. One reason that funds that perform poorly continue to perform poorly is that they have high expense ratios. Because mutual funds cannot be sold short, there is no way an investor can take advantage of this, except to withdraw money from poor-performing funds.[16]

Table 26.7 Persistence

		Measure Used			
		Ranking Measure	Evaluation Measure	Result	Positive Alpha for Top Group
1.	Grinblatt and Titman (1992)	G&T Measure	G&T Measure	Persistence	NR
2.	Hendricks, Patel, and Zeckhauser (1993)	Returns	Returns	Persistence	NR
3.	Brown and Goetzmarin (1995)	Returns	Returns CAPM 3-factor	Persistence Primarily worst group	NR
4.	Carhart (1997)	Returns	4-factor alpha	Lowest Decile	No
5.	Carhart (1997)	Alpha	4-factor alpha	Lowest & highest decile	Yes
6.	Elton, Gruber, and Blake (1996b)	Alpha	Alpha	Persistence	Yes
7.	Gruber (1996)	Alpha	Alpha	Persistence	Yes
8.	Cohen, Coval, and Pastor (2005)	Alpha	Alpha	Persistence	Yes
9.	Busse and Irvine (2006)	Bayesian Alpha	Alpha	Prediction	Yes
10.	Elton, Grober, and Blake (2011a)	Alpha	Alpha	Persistence	Yes
11.	Elton, Gruber, and Blake (2011d)	Alpha	Alpha	Persistence	Yes

NR means not relevant since the authors do not measure performance relative to index or set of indexes.

[15]The data from Elton, Gruber, and Blake (2013) have been annualized for presentation in Table 26.8. The results in each year report annual predictability over the period 1999–2009 for an average of mutual funds.
[16]See Gruber (1996) and Elton, Gruber, and Blake (1996a).

Table 26.8 Realized Alphas with Forecast in Previous Year (Work Data)

Rank in Previous Year	Realized Alpha
1	−0.048
2	−0.027
3	−0.020
4	−0.021
5	−0.016
6	−0.015
7	−0.011
8	−0.003
9	0.004
10	0.030
Spearman Corr.	**0.988**
P-Value	**<0.0001**

The results on well-performing funds are more mixed. While a few studies fail to find persistence among the best-performing funds, most do find persistence. For example, the top 10% of funds in Table 26.8 showed an average alpha in the following year of more than 1.5%. Furthermore, Elton, Gruber, and Blake (2012) employ a simulation study to show that the probability of this number arising by chance is less than 1/1000 of 1%.[17]

A theoretical argument against predictability is presented in Berk and Green (2004). Berk and Green argue that performance decreases with size, either because of increased costs and/or the need to accept less profitable investments. Because fund flows follow performance, investment will flow into any fund until performance above indexes is eliminated. Whether fund flows eliminate persistence depends on the amount of cash flows and how much and how quickly cost increases with size or how much and how quickly performance decreases with size. There have been four suggestions for why costs might increase or performance decrease as fund size grows: increasing fees, adding investments that are less promising to the portfolio (or indexing part of it), organization diseconomies, and transaction costs.

Expense ratios have two components: administrative costs (including sales costs) and management fees. For most funds the management fee schedule specifies that management fees will decrease with fund size in a particular manner. Changing the fee schedule is difficult and rarely done. Administrative costs have a large fixed component. Thus total fees as a percentage of assets decline with the size of the fund, and the relationship of expense ratios to size generally leads to performance increasing with size rather than decreasing.[18]

In addition, Pollet and Wilson (2008) have shown that as a fund grows larger, the number of securities changes only slightly. Thus if there are diseconomies of scale, they most likely involve transaction costs. These are being studied by a number of authors currently and should shed light on how long persistence should last.[19] The most relevant is Christoffersen, Keim, and Musto (2007). They studied Canadian mutual funds where trades have to be reported. They find that larger mutual funds have lower costs than smaller funds and that

[17]Investor awareness of this means that they will withdraw investments from bad-performing funds over time and account for the disappearance of funds from the industry (see Elton, Gruber, and Blake, 2012).

[18]Elton, Gruber, and Blake (2013) provide evidence that mutual funds that are successful and grow decrease fees.

[19]See, for example, Edelen, Evans, and Kadlec (2009) and Yan (2008).

active funds have lower trading costs than passive funds. They argue the latter is likely due to bunching of trades around index changes being more costly than the trading costs caused by active managers trading on information.[20]

Fama and French (2010) provide the first direct test of Berk and Green. Fama and French (2010) point out that the Berk and Green prediction that most fund managers have sufficient skill to cover their costs is not supported by the data. They examine the cumulative distribution of net returns using bootstrap simulation and conclude that bad-performing funds have risk-adjusted returns that are extremely unlikely to have arisen by chance, while those funds that have done extremely well may have obtained these results by chance. They do find that in the upper trail of performance there appear to be some funds that exhibited superior performance at a statistically significant level. Chen, Hong, Huang, and Kubik (2004) find that performance decreases with size and attribute this to organization diseconomies. However, despite this, they find predictability of performance. In addition, Elton, Gruber, and Blake (2012) find direct evidence that while persistence is weaker in very large funds, it still exists at levels that are statistically and economically meaningful.

Elton, Gruber, and Blake find that contrary to Berk and Green, expenses and management fees are smaller for well-performing funds and grow more slowly over time than the expense of bad-performing funds. However, Elton, Gruber, and Blake do find that performance does deteriorate over time as a function of cash flows. However, they present evidence that the erosion of performance as a function of past performance takes place slowly over time and that past performance does predict future performance for periods up to three years. Berk and Green have identified several factors that erode performance over time. They have misestimated the speed with which these influences impact performance. Past returns predict future returns even for the top-performing funds for periods in excess of two years.

Several studies have employed characteristics in addition to past return and expenses to predict performance. Kacpercyk, Salen, and Hend (2008) find that the gap between the return on a fund and the returns on a fund based on performance last period helps predict return. Cremers and Potajisto (2009) show that the difference between a fund's holdings and the holdings of its benchmarks can predict performance. Chevalier and Ellison (1999) look at the relationship between performance and manager characteristics such as age, time in the job, whether the manager had an MBA, and the average SAT scores at the schools they graduated from. Only the last had any predictive power.

Another question to ask is whether funds that have higher alpha have higher future cash flow. Since alpha predicts future alpha, we would expect some investors to be aware of this and invest in good-performings. The fact that investors do so and that marginal flows into mutual funds outperform indexes is documented by Gruber (1996) and Zheng (1999). Sirri and Tufano (1998) present evidence that cash flow is negative related to expense ratios and volatility.

Performance in the Hedge and Commodity Fund Industries

Equilibrium models together with the assumption of a nearly efficient public market for securities suggest that the greater part of an investor's expected return is generated by exposure to systematic risk factors. There are some sectors of the investment industry, however, that focus on generating returns solely through manager skill and often have low

[20]Keirn and Madhaven (1995, 1997) find that execution size increases transaction costs for institutional traders using plexus data. However, they cannot tell if the smaller orders were simply a bigger order being executed as a series of small orders or a small order.

exposure to systematic factors. While active trading partnerships predicated on managerial ability have undoubtedly existed for as long as there has been speculation in the securities markets, recently there has been a burst of activity in this section. The hedge fund industry in the United States has grown from a handful of firms in the 1940s, 1950s, and 1960s to a universe of thousands of managers and trillions of dollars.

It is very difficult to generalize about this vast industry; however, most hedge funds share a few features in common. First, they seek returns through manager acumen—whether through the application of complex statistical models or through fundamental research. Second, for the most part, hedge fund managers invest in publicly traded securities, as opposed to private equity managers who hold positions in nonpublicly traded firms. Third, the compensation structure for hedge fund managers typically has two components—a fixed fee of 1% or 2% of assets each year plus an incentive fee equal to a percentage of the return on assets each year that is paid if the manager generates a return in excess of an agreed-upon benchmark. This incentive fee has often had the additional feature that losses from previous years must be made up before the incentive fee is payable in the current year. This last characteristic is called a *high water mark* feature.

A closely related asset class is commodity pools. These traditionally are actively managed funds like hedge funds, but they invest in commodity futures rather than the entire spectrum of public securities. Finally, a special class of hedge funds is called a "fund-of-fund," which does not invest directly in individual securities but instead invests in the shares of other hedge funds and commodity funds. They presumably provide three services: selection of managers, monitoring of risk, and access to managers who are not accessible to the general investor.

Taken as a group, hedge funds, commodity funds, and funds-of-funds are largely predicated on skill. Indeed, the presumption of the existence of manager skill would naturally lead us to look at this universe to examine whether managers are able to beat a benchmark. The reason for this is the prevalence of incentive fees. The typical incentive fee in the hedge fund industry is 20% of the profits annually. If a manager believes in his ability to regularly beat a benchmark, the hedge fund industry is the ideal industry in which to operate—passive investors, in effect, are financing managerial talent.

Strangely enough, the evidence on the existence of skill in the hedge fund and commodity fund industry is mixed at best. Elton, Gruber, and Rentzler (1990) studied commodity fund performance over the period 1980 through 1988 and found that, on average, they provided returns below T-bill rates with high risk. Brown, Goetzmann, and Ibbotson (1999) studied the hedge fund industry through the period 1989–1995, using a survivorship-bias-free database. They found that, as a whole, a major sector of hedge funds—offshore funds—provided superior risk-adjusted returns over a seven-year period. However, classical tests of performance persistence were disappointing. Oddly enough, in an industry predicated on manager skill, the winners in the industry did not tend to repeat. This negative evidence was not affected by classification within different styles or management's focus on different security types. Later studies, using longer time series, have changed these results somewhat. Agarwal and Naik (2000) found quarterly persistence of hedge fund returns, and some of their other work with coauthors suggests that compensation affects performance.[21] Few studies have found any positive evidence with regard to fund-of-fund performance.[22]

[21]Agarwal and Naik (2000).

[22]Fung and Hsieh (2005).

Special Issues with Hedge Funds

As a general approach, hedge funds earn returns by providing liquidity to markets by engaging in trading activity considered inappropriate for investors lacking the knowledge or resources to do so. They engage in short sales, leveraged transactions, and other strategies not available to mutual funds and pension funds that are more heavily regulated in the interests of their investors and stakeholders.[23] This goal leads to some important issues with respect to performance monitoring and measurement. First, hedge funds are likely to trade in relatively inefficient markets, for example, small, illiquid securities, poorly understood derivative instruments, or securities of companies in distress or involved in mergers. In seeking to buy undervalued securities, they are likely to be investing in assets that at any given point in time are out of favor with most other investors. This occasionally makes the pricing of hedge fund portfolios problematic. Chan, Getmansky, Hass, and Lo (2005) study the time series characteristics of hedge funds and find positive autocorrelation—evidence suggesting that some securities in the portfolio may be priced using old prices. They note that this tends to artificially smooth hedge fund return and to potentially misrepresent the volatility of the investment.

Fung and Hsieh (2001) point out that hedge fund returns are highly nonlinear—that is, they display option-like characteristics with respect to standard benchmarks. This is not surprising in light of the fact that hedge fund managers may seek to add value through timing as well as through security selection. They find that adjusting for risk exposure when these exposures are changing through time requires sophisticated modeling and lots of data.

Brown, Goetzmann, and Park (2001) used weekly data for a set of "macro" hedge funds that were active in trading international currency futures and non-U.S. securities around the time of the Asian currency crisis in 1997. They found that fund exposures to currencies could and did change rapidly over intervals shorter than one month, as these managers tried to trade on the evolving international sentiments about Asian markets. A few recent studies have had access to individual trade data by hedge funds. Griffin, Harris, and Topaloglu (2002) and Busse and Nagle (2004) study hedge fund trades around 2000, after U.S. technology stocks had risen to all-time highs. They both find that hedge funds tended to be "momentum" traders, making money by actively speculating on short-term positive trends in securities and changing their holdings with high frequency. As discussed earlier, high-frequency trading can present special challenges for performance measurement. High-frequency trading also leads to a need for performance monitoring—that is, the evaluation of the risks as well as return. Consider the following hypothetical problem. A trader is almost certainly going to be dismissed if the fund return over the next month is less than −20%. By the same token, the trader will not earn an incentive fee unless the returns are positive. Now, suppose the trader can trade an infinite number of times before the month is over and the investor cannot observe the trader's position until the month is over. In this circumstance, the trader can generate a probability of a positive return of close to 1. If the fund takes a loss on the first day, it is only necessary to "double" the bet on the following day. If that bet is successful, the loss is erased. If it fails, then in the next instance, the trader may again double up. As long as the trader has enough credit or cash to keep doubling, it is possible to eliminate any past loss. Of course, a string of five or six losses in a row exponentially increases the first loss, and any interim evaluation of the fund portfolio might reveal an extreme position or even a negative value. This is not a free lunch; while

[23]The term *hedge fund* is not descriptively accurate. There is no such thing as a well-defined hedge fund strategy or approach to investing, and hedge funds are not generally "hedged" in any meaningful sense. Rather, in the United States, a hedge fund is a limited investment partnership otherwise exempt from registering with the Securities and Exchange Commission under Section 3C1 and 3C7 of the Investment Company Act of 1940. As Brown and Goetzmann (2003) show, there is a remarkable diversity of styles of management under the hedge fund banner.

the probability of having a loss at the end of the month is small, when a loss does occur, it will be gigantic. Brown et al. (2005) developed a test for such doubling strategies and applied it to daily returns of some Australian money managers. They found some evidence that managers occasionally use this risky strategy. Doubling is, of course, not the only strategy a fund manager might use to adjust the position of the fund between accountings. Goetzmann et al. (2005) document a range of "informationless" dynamic techniques that can be used to "game" standard performance measures such as Sharpe ratios, multifactor alpha measures, and timing measures.

Transparency By providing liquidity to otherwise illiquid markets, hedge funds seek to exploit market inefficiencies. To the extent that these techniques successfully generate risk-adjusted returns, the industry argues that they represent a comparative advantage that could be lost if the techniques were common knowledge. As a result, because hedge funds in the United States are not required to submit public disclosure statements, portfolio disclosure and reporting of hedge fund security positions is rare. The resulting lack of transparency puts a particular burden on investors to conduct appropriate due diligence. An analysis of hedge fund flows by Brown et al. (2008a) shows that while sophisticated investors are aware of operational risks associated with conflicts of interest and potential capital loss, most hedge fund investors either do not have this information or regard it as immaterial. Brown et al. (2008a) argue that by weeding out funds with high degrees of operational risks, appropriate due diligence can generate significant risk-adjusted returns in a diversified hedge fund portfolio strategy.

APPENDIX

The Use of APT Models to Evaluate and Diagnose Performance

The use of an APT model in combination with a multi-index model allows for better diagnoses of what a portfolio manager is doing, a better development of appropriate benchmarks, and a better measurement and attribution of performance.[24]

The overall performance and reasons for the performance can be measured using an APT model such as the one discussed in Chapter 16. As an example, consider the model we described in the latter section of Chapter 16. Let us examine the return-generating process presented in Equation (16.11) with the APT model and the associated λ values presented in Chapter 16 inserted:

$$\begin{aligned} R_i - R_F = &-4.32b_{iI} + 1.49b_{iS} + 0.00b_{iO} + 3.96b_{iM} \\ &+ b_{iI}I_I + b_{iS}I_S + b_{iO}I_O + b_{iM}I_M \end{aligned} \tag{26-A-1}$$

Recall that the subscript

I stands for inflation

S stands for aggregate sales

O stands for oil prices

M stands for the S&P index with other influences removed

Now assume a particular manager *x* had the following values for the sensitivities (*b*s) on the portfolio:

$$b_{xI} = -0.5 \quad b_{xS} = 2.75 \quad b_{xO} = -1.00 \quad b_{xM} = 1.30$$

[24]There are many APT models. Whether using an APT model or the multi-index models of the prior section results in better performance evaluation depends on how well the model that is used approximates the return-generating process or the true APT model.

Assume the average *b*s for the S&P index were

$$b_{\text{S\&P }I} = -0.37 \quad b_{\text{S\&P }S} = 1.71 \quad b_{\text{S\&P }O} = 0 \quad b_{\text{S\&P }M} = 1.0$$

Let us further assume that the difference of each index from its expected value in the period where we are evaluating the manager is

$$I_I = 0.7 \quad I_S = 0.5 \quad I_O = 0.4 \quad I_M = 1.00$$

Assume that the manager is free to select the sensitivities to each index and that we want to investigate why the manager does better or worse than the S&P index. Further assume that the manager was able to earn 14.52% excess return (above the riskless rate of interest). We can decompose the manager's return (compared to the S&P index) into the following categories: expected return from the S&P index, extra return on the S&P index that is earned from factors having returns that are different from their expected value, extra expected return earned from having sensitivities different from those on the S&P index, extra return earned from having sensitivities different from those of the S&P index that is earned because factors have return that are different from their expected values, and extra return from security selection.

The expected return from the S&P index is simply from Equation (26A-1), recognizing that all *I*s have an expected value of zero:

$$\bar{R}_{\text{S\&P}} - R_F = -4.32\beta_{\text{S\&P }I} + 1.49\beta_{\text{S\&P }S} + 0.00\beta_{\text{S\&P }O} + 3.96\beta_{\text{S\&P }M}$$
$$\bar{R}_{\text{S\&P}} - R_F = -4.32(-0.37) + 1.49(1.71) + 0.00(0.0) + 3.96(1) = 8.103$$

Now, in any period, the excess return on the S&P index will differ from this because the returns on the factors over the period are not at their expected values (I_js $\neq$ 0). To see the influence of this, simply multiply the betas for the S&P times the value that the I_js take on over the evaluation period, or

$$\left(R_{\text{S\&P}} - \bar{R}_{\text{S\&P}}\right) = \beta_{\text{S\&P }I} I_I + \beta_{\text{S\&P }S} I_S + \beta_{\text{S\&P }O} I_O + \beta_{\text{S\&P }M} I_M$$
$$\left(R_{\text{S\&P}} - \bar{R}_{\text{S\&P}}\right) = (-0.37)(0.7) + (1.71)(0.5) + (0.0)(0.4) + (1.0)(1.0) = 1.591$$

To this point we have seen that the excess return on the S&P was 9.694 with a return of 8.103 expected and 1.591 due to the fact that the factors driving security returns had returns different from expected returns over the period.

The choice of different sensitivities impacts performance in two ways. First, with different sensitivities, the return for risk bearing (expected return) will differ. Second, different sensitivities will affect the additional return that may be gained or lost because the return on an index was different from that required by the average investor (nonzero *I*s).

Let us think about this for a moment. The manager discussed earlier chose to have a higher sensitivity to the residual market influence than the S&P index had. Because the market price of this risk is positive, we would expect a manager with higher sensitivity to earn on average a higher return. However, this increase in expected return is simply what investors require for the extra risk. If we were to credit this extra return to the manager's performance, all managers who were not constrained would on average hold portfolios with higher betas with all factors that had positive λs.

The extra return required (expected) by investors because this manager has chosen to take extra risk is simply the difference in sensitivity between the portfolio and the S&P index times the associated λ. This is shown in Table 25.10 under the column entitled "Differential Expected Return."

Table 26A-1 Effect of Different Sensitivities on Performance

Common Influences	a Sensitivity of Manager's Portfolio b_{xj}	b Sensitivity of Market Portfolio $b_{S\&Pj}$	c Differential Sensitivity $c = a - b$	d Expected Return on Influence λ	e Unexpected Return on Influence l_j	f Differential *Expected* Return $f = c \times d$	g Differential *Unexpected* Return $g = c \times e$
Inflation	−0.5	−0.37	−0.13	−4.32	0.7	0.512	−0.091
Sales growth	2.75	1.71	1.04	1.49	0.5	1.550	0.520
Oil prices	−1.00	0.0	−1.00	0.00	0.4	0.000	−0.400
Market	1.30	1.0	0.30	3.96	1.0	1.188	0.300
						3.300	0.329

We now come to the examination of any payoff that the investor receives due to the manager's ability to appropriately adjust factor sensitivities. This is the product of the difference between the factor sensitivities the manager chooses and the S&P index and the return on any factor that was not required (expected) as compensation for risk. It is the sum of the differential b_{ij}s times the I_js and is shown in the column entitled "Differential Unexpected Return" in Table 26A-1.

Adding together these four elements of return, we get 13.323%. Because the manager was able to earn an excess return of 14.52%, the difference of 1.197% is due to security selection. These ideas are summarized in Table 26A-2.

Although this type of decomposition of performance is extremely useful, one must be careful in deciding what to attribute to management skill. How much of the return should be attributed to management skill? Clearly the 1.19 for security selection is attributed to management skill. However, the issue that needs to be addressed is whether the manager should be given credit for the return earned by having sensitivities different from the S&P index and having indexes having returns different from expected (0.329%). If the manager is free to choose the sensitivities, then one can argue that the extra return from a good choice of sensitivities should be attributed to management skill.

Note, however, that the factor sensitivities for this manager exactly match those of the average growth stock manager discussed in Chapter 16. If this manager was hired to act as a typical growth stock manager, then the extra return due to sensitivities differing from the S&P did not result from an active choice on the part of the manager. In this case the only extra return she can be credited for is the 1.197% due to selectivity.

In general, this raises the question of any appropriate benchmark to use in evaluating the manager. We have used the S&P index in our example and showed how the results would change had we used an average growth stock manager as the benchmark. In general, there are two types of benchmarks that seem appropriate. The first is a bogey that the investor selects as the target the manager is asked to outperform. This bogey can be an index, a portfolio, or another manager or group of managers. If this type of bogey is selected, then all of the analysis we have done previously holds, except that the *b* values for the bogey are used wherever we used *b* values for the S&P index.

The second type of bogey is the average *b* for the manager under question. If the manager is judged against the average beta, then all excess returns that arise from differences from this average should be attributed to management performance.

Table 26A-2 Decomposition of Performance Using APT

Return on Benchmark	
a. Expected	8.103
b. From Factors Deviating from Mean	1.591
Return from Different Sensitivities (*b*)	
a. Expected	3.300
b. From Factors Deviating from Mean	0.329
Return from Security Selection	1.197
Total Return on Fund	14.52

QUESTIONS AND PROBLEMS

1. Here are data on five mutual funds:

Fund	Return	Standard Deviation	Beta
A	14	6	1.5
B	12	4	0.5
C	16	8	1.0
D	10	6	0.5
E	20	10	2

What is the reward-to-variability ratio and the ranking if the risk-free rate is 3%?

2. For the data in Problem 1, what is the Treynor measure and ranking?

3. For the data in Problem 1, what is the differential return if the market return is 13%, the standard deviation of return is 5%, and standard deviation is the appropriate measure of risk?

4. For the data in Problem 1, what is the differential return if beta is the appropriate measure of risk?

5. Assume that the zero-beta form of the capital asset pricing model (CAPM) is appropriate. What is the differential return for the funds shown in Problem 1 if $R_z = 4\%$?

6. For funds *A* and *B* in Problem 1, how much would the return on B have to change to reverse the ranking using the reward-to-variability measure?

BIBLIOGRAPHY

1. Admati, Anat, Phleiderer, Paul, Pers, Stephen, and Bhattacharya, Sudipto. "On Timing and Selectivity," *Journal of Finance*, **41** (1986), pp. 715–730.
2. Agarwal, Vikas and Naik, Narayan Y. "Multi-Period Performance Persistence Analysis of Hedge Funds," *Journal of Finance and Quantitative Analysis*, **35** (2000), pp. 327–342.
3. Baks, Klaas P., Metrick, Andrew, and Wachter, Jessica. "Should Investors Avoid All Actively Managed Mutual Funds: A Study in Bayesian Performance Evaluation," *Journal of Finance*, **56** (2001), pp. 45–85.
4. Baker, Malcolm, Litov, Lubomir, Wachter, Jessica, and Wurgler, Jeffrey. "Can Mutual Fund Managers Pick Stocks? Evidence from Their Trades Prior to Earnings Announcements," *Journal of Financial and Quantitative Analysis*, **45** (2004), pp. 1111–1131.
5. Ball, Ray, Kothari, S. P., and Shanken, Jay. "Problems in Measuring Portfolio Performance: An Application to Contrarian Investment Strategies," *Journal of Financial Economics*, **38** (1995), pp. 79–107.
6. Barber, Brad, Odean, Terrence, and Zheng, Lu. "Out of Sight Out of Mind: Effects of Expense on Mutual Fund Flows," *Journal of Business*, **78** (2005), pp. 2095–2119.
7. Becker, Connie, Ferson, Wayne, Myers, David, and Schill, Michael. "Conditional Market Timing with Benchmark Investors," *Journal of Financial Economics*, **52** (1999), pp. 119–148.
8. Bergstresser, Daniel, and Poterba, James. "Do After-Tax Returns Affect Mutual Fund Inflows?" *Journal of Financial Economics*, **63** (2002), pp. 381–414.
9. Berk, Jonathan, and Green, Richard. "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, **112** (2004), pp. 1269–1295.
10. Berk, Jonathan, and Stanton, Richard. "Managerial Ability, Compensation and the Closed End Fund Discount," *Journal of Finance*, **62** (2007), pp. 529–556.
11. Blake, Christopher R., Elton, Edwin J., and Gruber, Martin J. "The Performance of Bond Mutual Funds," *Journal of Business*, **66** (1994), pp. 371–403.
12. Bollen, Nicolas P. B., and Busse, Jeffrey A. "On the Timing Ability of Mutual Fund Managers," *Journal of Finance*, **56** (2001), pp. 1075–1094.

13. Brown, Keith, Harlow, Van, and Starks, Laura. "Incentives in the Mutual Fund Industry," *Journal of Finance*, **5** (1996), pp. 85–110.
14. Brown, Stephen J., Gallagher, David R., Steenbeck, Onno W, and Swan, Peter I. "Doubler or Nothing: Patterns of Equity Fund Holdings and Transactions," EFA Moscow Meeting Paper, (2005).
15. Brown, Stephen J., Goetzman, William N., and Ibbotson, Roger G. "Offshore Hedge Funds: Survival and Performance 1989–1995," *Journal of Business*, **72** (1999), pp. 91–117.
16. Brown, Stephen J., Goetzmann, William N., Ibbotson, Roger G., and Ross, Stephen A. "Survivorship Bias in Performance Studies," *Review of Financial Studies*, **5** (1992), pp. 553–580.
17. Brown, Stephen J., Goetzman, William N., and Park, James M. "Conditions for Survival: Changing Risk and the Performance of Hedge Fund Managers and CTAS," *Journal of Finance*, **61** CTAS (2001), pp. 1869–1886.
18. Brown, Stephen, and Goetzmann, William. "Hedge Funds with Style," *Journal of Portfolio Management*, **29** (2003), pp. 101–112.
19. Brown, Stephen, Goetzmann, William, Liang, Bing, and Schwartz, Christopher. "Mandatory Disclosure and Operational Risks: Evidence from Hedge Fund Registration," *Journal of Finance*, **63** (2009), pp. 2786–2815.
20. Brown, Stephen, Fraser, Thomas, and Liang, Bing. "Hedge Fund Diligence: A Source of Alpha in a Hedge Fund Portfolio," *Journal of Investment Management*, **26** (2008), pp. 23–33.
21. Brown, Stephen J., and Goetzmann, William N. "Performance Persistence," *Journal of Finance*, **50** (1995), pp. 679–698.
22. Brown, Stephen J., and Goetzmann, William N. "Mutual Fund Styles," *Journal of Financial Economics*, **43** (1997), pp. 373-399.
23. Busse, Jeffrey. "Volatility Timing in Mutual Funds: Evidence from Daily Returns," *Review of Financial Studies*, **12** (1999), pp. 1009–1041.
24. Busse, Jeffrey A., and Irvine, Paul J. "Bayesian Alphas and Mutual Fund Persistence," *Journal of Finance*, **61** (2006), pp. 2251–2288.
25. Carhart, Mark M. "On Persistence in Mutual Fund Performance," *Journal of Finance*, **52** (1997), pp. 57–82.
26. Carpenter, Jennifer, and Lynch, Anthony. "Survivorship Bias and Attrition Effects in Measures of Performance Persistence," *Journal of Financial Economics*, **54** (1999), pp. 337–374.
27. Chan, Louis, Chen, Hsui, and Lakonishok, Josef. "On Mutual Fund Investment Styles," *Review of Financial Studies*, **15** (2002), pp. 1407–1437.
28. Chan, Louis, Dimmock, Stephen, and Lakinishok, Josef. "Benchmarking Money Manager Performance: Issues and Evidence," *Review of Financial Studies*, **22** (2010), pp. 4553–4599.
29. Chan, Nicholas, Getmansky, Mila, Hass, Shane M., and Lo, Andrew W. "Systematic Risk and Hedge Funds." In M. Carey and R. Stulz (eds.), *The Risks of Financial Institutions and the Financial Sector* (Chicago: University of Chicago Press, 2005).
30. Chance, Don, and Hemler, M. C. "Performance of Professional Market Timers: Data Evidence from Executed Strategies," *Journal of Financial Economics*, **62** (2001), pp. 377–411.
31. Chang, E., and Lewellen, W. "Market Timing and Mutual Fund Investment Performance," *Journal of Business*, **57** (1983), pp. 57–72.
32. Chen, Joseph, Hong, Harrison, Huang, Ming, and Kubik, Jeffrey. "Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization," *American Economic Review*, **94** (2004), pp. 1276–1307.
33. Chen, Wai-Fu, Roll, Richard, and Ross, Steve. "Economic Forces and the Stock Market," *Journal of Business*, **56** (1986), pp. 383–403.
34. Chen, Yong, Ferson, Wayne, and Peters, Helen. "Measuring the Timing Ability and Performance of Bond Mutual Funds," *Journal of Financial Economics*, **98** (2010), pp. 72–89.
35. Chen, Zhiwu, and Knez, Peter. "Portfolio Performance Measurement: Theory and Application," *Review of Financial Studies*, **9** (1996), pp. 511–555.
36. Chevalier, Judith, and Ellison, Glenn. "Risk-Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, **105** (1997), pp. 1167–1200.
37. Chevalier, Judith, and Ellison, Glenn. "Are Some Mutual Fund Managers Better than Others: Cross-Sectional Patterns in Behavior and Performance," *Journal of Finance*, **54** (1999), pp. 875–900.

38. Christoffersen, Susan, and Musto, David. "Demand Curves and the Pricing of Money Management," *Review of Financial Studies*, **15** (2002), pp. 1499–1524.
39. Christoffersen, Susan, Keim, Donald, and Musto, David. "Valuable Information and Costly Liquidity: Evidence from Individual Mutual Fund Trades," unpublished manuscript, University of Pennsylvania (2007).
40. Christopherson, Jon, Ferson, Wayne, and Glassman, Debra. "Conditioning Manager Alpha on Economic Information: Another Look at the Persistence of Performance," *Review of Financial Studies*, **11** (1998), pp. 111–142.
41. Cohen, R., Coval, J., and Pastor, L. "Judging Fund Managers by the Company They Keep," *Journal of Finance*, **60** (2005), pp. 1057–1096.
42. Comer, George. "Hybrid Mutual Funds and Market Timing Performance," *Journal of Business*, **79** (2005), pp. 771–797.
43. Comer, George, and Rodriguez, Javier. "Corporate and Government Bond Funds: An Analysis of Investment Style, Performance and Cash Flow," working paper, Georgetown University (2006).
44. Connor, Gregory, and Korajczyk, Robert A. "Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis," *Journal of Financial Economics*, **15** (1986), pp. 373–394.
45. Connor, Gregory, and Korajczyk, Robert A. "Risk and Return in an Equilibrium APT: Applications of a New Test Methodology," *Journal of Financial Economics*, **21** (1988), pp. 255–289.
46. Connor, Gregory, and Korajczyk, Robert. "The Attributes Behavior and Performance of U.S. Mutual Funds," *Review of Quantitative Finance and Accounting*, **1** (1991), pp. 4–25.
47. Cooper, Michael, Gulen, Huseyin, and Rau, Raghavendra. "Changing Names with Style: Mutual Fund Name Changes and Their Effects on Fund Flows," *Journal of Finance*, **60** (2005), pp. 2825–2858.
48. Cornell, Brad. "Asymmetric Information and Portfolio Performance Measurement," *Journal of Financial Economics*, **7** (1979), pp. 381–390.
49. Coval, Joshua D., Hirshleifer, David A., and Shumway, Tyler. "Can Individual Investors Beat the Market?" NOM working paper, Harvard University (2003).
50. Cremers, Martijn, and Petajisto, Antti. "How Active Is Your Fund Manager? A New Measure that Predicts Performance," *Review of Financial Studies*, **22** (2009), pp. 3329–3365.
51. Cremers, Martijn, Petajisto, Antti, and Zitzewitz, Eric. "Should Benchmark Indices Have Alpha? Revisited Performance Evaluation," working paper, Yale University (2010).
52. Crenshaw, T. E. "Evaluation of Investment Performance," *Journal of Business*, **50** (1977), pp. 462–485.
53. Dahlquist, Magnus, and Soderlind, Paul. "Evaluating Portfolio Performance with Stochastic Discount Factors," *Journal of Business*, **72** (1999), pp. 347–384.
54. Daniel, Kent, Grinblatt, Mark, Titman, Sheridan, and Wenners, Russ. "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks," *Journal of Finance*, **52** (1997), pp. 1035–1058.
55. Del Guercio, Diane, and Tkac, Paula A. "The Determinants of the Flow of Funds of Managed Portfolios: Mutual Funds v. Pension Funds," *Journal of Financial and Quantitative Analysis*, **37**, (2002), pp. 523–557.
56. Del Guercio, Diane, and Tkac, Paula A. "Star Power: The Effect of Morningstar Ratings on Mutual Fund Flow," *Journal of Financial and Quantitative Analysis*, **43** (2008), pp. 907–936.
57. Duffee, Gregory R. "Tenn Premia and Interest Rate Forecasts in Affine Models," *Journal of Finance*, **57** (2002), pp. 405–443.
58. Edelen, Roger, Evans, Richard, and Kadlec, Gregory. "Scale Effects in Mutual Fund Perfor-Mance: The Role of Trading Costs," unpublished manuscript, University of Virginia (2009).
59. Edwards, Amy, Harris, Lawrence, and Piwowar, Michael S. "Corporate Bond Market Transaction Costs and Transparency," *Journal of Finance*, **61** (2006), pp. 1361–1397.
60. Elton, Edwin J., Gruber, Martin J., Das, Sanjiv, and Hlavka, Matthew. "Efficiency with Costly Information: A Reinterpretation of Evidence from Manager Portfolios," *Review of Financial Studies*, **6** (1993), pp. 1–23.
61. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Fundamental Variables, APT, and Bond Fund Performance," *Journal of Finance*, **50** (1995), pp. 1229–1256.

62. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher. "Survivorship Bias and Mutual Fund Performance," *Review of Financial Studies*, **9** (1996a), pp. 1097–1120.
63. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "The Persistence of Risk-Adjusted Mutual Fund Performance," *Journal of Business*, **69** (1996b), pp. 133–157.
64. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Common Factors in Active and Passive Portfolios," *European Finance Review*, **3** (1999), pp. 53–78.
65. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "A First Look at the Accuracy of the CRSP Mutual Fund Database and a Comparison of the CRSP and Morningstar Mutual Fund Databases," *Journal of Finance*, **56** (2001), pp. 2415–2450.
66. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Incentive Fees and Mutual Funds," *Journal of Finance*, **58** (2003), pp. 779–804.
67. Elton, Edwin J., Gruber, Martin J., and Busse, Jeff. "Are Investors Rational? Choices Among Index Funds," *Journal of Finance*, **58** (2004), pp. 427–465.
68. Elton, Edwin J., Gruber, Martin J., Brown, Stephen J., and Goetzmann, William. *Modem Portfolio Theory and Investment Analysis*, 8th ed. (Hoboken, NJ: John Wiley, 2010a).
69. Elton, Edwin J., Gruber, Martin J., Blake, Christopher R., Krasny, Yoel, and Ozelge, Sadi. "The Effect of the Frequency of Holdings Data on Conclusions about Mutual Fund Management Behavior," *Journal of Banking and Finance*, **34** (2010b), pp. 912–922.
70. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Holdings Data, Security Returns, and the Selection of Superior Mutual Funds," *Journal of Financial and Quantitative Analysis*, **46,** No. 2 (April 2011), pp. 341–367.
71. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "An Examination of Mutual Fund Timing Using Monthly Holdings Data," *Journal of Banking and Finance*, **4** (2011b), pp. 912–922.
72. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Why Do Closed End Funds Exist? An Additional Explanation for the Growth in Domestic Closed End Bond Funds," unpublished manuscript, New York University (2011c).
73. Elton, Edwin J., Gruber, Martin J., and Blake, Christopher R. "Does Size Matter? The Relationship Between Size and Performance," *Review of Asset Pricing Studies*, **2** (2012), pp. 245–74.
74. Fama, Eugene F., and French, Ken R. "Multifactor Explanations of Asset Pricing Anomalies," *Journal of Finance*, **51** (1996), pp. 55–87.
75. Fama, Eugene F., and French, Ken R. "Luck versus Skill in the Cross-Section of Mutual Fund Return," *Journal of Finance*, **65** (2010), pp. 1915–1947.
76. Farnsworth, Heber, Ferson, Wayne, Jackson, David, and Todd, Steven. "Performance Evaluation with Stochastic Discount Factors," *Journal of Business*, **75** (2000), pp. 473–504.
77. Ferson, Wayne E., and Schadt, Rudi W. "Measuring Fund Strategy and Performance in Changing Economic Conditions," *Journal of Finance*, **51** (1996), pp. 425–461.
78. Ferson, Wayne E., and Khang, Kenneth. "Conditional Performance Measurement Using Portfolio Weights; Evidence for Pension Funds," *Journal of Financial Economics*, **65** (2002), pp. 249–282.
79. Ferson, Wayne E., and Merijun, Qian. "When Can Market Timers Time?" unpublished manuscript, University of Southern California (2006).
80. Friend, Irwin, Blume, Marshall, and Crocket, Gene. "Measurement of Portfolio Performance under Uncertainty," *American Economic Review*, **60** (1970), pp. 561–575.
81. Fung, William, and Hsieh, David. "The Risk in Hedge Fund Strategies: Theory and Evidence from Trend Followers," *Review of Financial Studies*, **14** (2001), pp. 313–341.
82. Gendron, Michel, and Genest, Christian. "Performance Measurement under Asymmetric Information and Investment Constraints," *Journal of Finance*, **45** (1990), pp. 1655–1661.
83. Glode, Vincent. "Why Mutual Funds Underperform," *Journal of Financial Economics*, **99** (2011), pp. 546–559.
84. Goetzmann, William N., and Ibbotson, Roger G. "Do Winners Repeat: Predicting Mutual Fund Performance," *Journal of Portfolio Management*, **20** (1994), pp. 9–18.
85. Goetzmann, William, Ingersoll, Jonathan Jr., Speigel, Matthew, and Welch, Ivo. "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures," *Review of Financial Studies*, **20** (2007), pp. 1503–1546.

86. Goetzmann, William N., Ingersoll, Jonathan E. Jr., and Ivkovich, Zoran. "Monthly Measurement of Daily Timers," *Journal of Financial and Quantitative Analysis*, **95** (2000), pp. 257–290.
87. Goetzmann, William N., and Kumar, Llok. "Equity Portfolio Di Versification," Yale School of Management working papers YSM 17, Yale School of Management (2004).
88. Goetzmann, William, and Peles, Nadav. "Cognitive Dissonance and Mutual Fund Investors," *Journal of Financial Research*, **20** (1997), pp. 145–158.
89. Graham, John, and Harvey, Campbell. "Market Timing Ability and Volatility Implied in Investment Newsletters' Asset Allocation Recommendations," *Journal of Financial Economics*, **42** (1996), pp. 397–422.
90. Grinblatt, Mark, and Titman, Sheridan. "Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings," *Journal of Business*, **62** (1989a), pp. 393–416.
91. Grinblatt, Mark, and Titman, Sheridan. "Portfolio Performance Evaluation: Old Issues and New Insights," *Review of Financial Studies*, **2** (1989b), pp. 393–422.
92. Grinblatt, Mark, and Titman, Sheridan. "The Persistence of Mutual Fund Performance," *Journal of Finance*, **47** (1992), pp. 1977–1984.
93. Grinblatt, Mark, and Titman, Sheridan. "Performance Measurement without Benchmarks: An Examination of Mutual Fund Returns," *Journal of Business*, **66** (1993), pp. 47–68.
94. Gruber, Martin J. "Another Puzzle: The Growth in Actively Managed Mutual Funds," *Journal of Finance*, **51** (1996), pp. 783–810.
96. Han, Son, and Li, Dan. "Liquidity Crisis Runs and Security Design," working paper, Federal Reserve Board (2009).
97. Harris, Lawrence, and Piwowar, Michael. "Municipal Bond Liquidity," *Journal of Finance*, **61** (2006), pp. 1361–1397.
98. Hasbrouck, Joel. "Intraday Price Formation in the Market for U.S. Equity Markets," *Journal of Finance*, **58** (2003), pp. 2375–2400.
99. Hendricks, Darryl, Patel, Jayendu, and Zeckhauser, Richard. "Hot Hands in Mutual Funds: Short-Run Persistence of Relative Performance, 1974–1988," *Journal of Finance*, **48** (1993), pp. 93–130.
100. Henriksson, Roy D., and Merton, Robert C. "On Market Timing and Investment Performance II: Statistical Procedures for Evaluating Forecasting Skills," *Journal of Business*, **54** (1981), pp. 513–534.
101. Henriksson, Roy D. "Market Timing and Mutual Funds Performance: An Empirical Investigation," *Journal of Business*, **57** (1984), pp. 73–96.
102. Investment Company Institute. *Investment Company Factbook* (Washington, DC: Investment Company Institute, 2011).
103. Ippolito, Richard A. "Efficiency with Costly Information: A Study of Mutual Fund Performance, 1965–1984," *Quarterly Journal of Economics*, **104** (1989), pp. 1–24.
105. Jagannathan, Ravi, and Korajczyk, Robert A. "Assessing the Market Timing Performance of Managed Portfolios," *Journal of Business*, **59** (1986), pp. 217–236.
106. Jegadeesh, Narasimham, and Titman, Sheridan. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance*, **48** (1993), pp. 93–130.
107. Jensen, Michael C. "The Performance of Mutual Funds in the Period 1945–1964," *Journal of Finance*, **23** (1968), pp. 89–416.
108. Jiang, George, Yao, Tong, and Yu, Tong. "Do Mutual Funds Time the Market? Evidence from Holdings Data," *Journal of Financial Economics*, **88** (2007), pp. 119–145.
109. Jobson, J. D., and Korkie, Bob. "Potential Performance and Tests of Portfolio Efficiency," *Journal of Financial Economics*, **10** (1982), pp. 443–466.
110. Jobson, J. D., and Korkie, Bob. "On the Jensen Measure and Marginal Improvements in Portfolio Performance; a Note," *Journal of Finance*, **39** (1984), pp. 245–252.
111. Kacperczyk, Marcin T., Sialm, Clemens, and Zheng, Lu. "Unobserved Actions of Mutual Funds," *Review of Financial Studies*, **21** (2008), pp. 2379–2416.
112. Kaniel, Ron, Sarr, Gideon, and Titman, Sheridan. "Individual Investor Sentiment and Stock Returns," working paper, University of Texas (2004).
113. Kaplan, Steven N., and Sensoy, Berk A. "Do Mutual Funds Time Their Benchmarks?" working paper, University of Chicago (2005).

114. Keirn, Donald B., and Madhavan, Ananth. "Empirical Evidence on the Behavior of Institutional Traders," *Journal of Financial Economics*, **37** (1995), pp. 371–399.
115. Keirn, Donald B., and Madhavan, Ananth. "Transaction Costs and Investment Style: An Interexchange Analysis of Institutional Equity Trades," *Journal of Financial Economics*, **46** (1997), pp. 265–292.
116. Kent, Daniel. "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks," *Journal of Finance*, **52**, pp. 1035–1058.
117. Kothari, S., and Warner, Jerold. "Evaluating Mutual Fund Performance," *Journal of Finance*, **56** (2001), pp. 1985–2010.
118. Lehmann, Bruce N., and Modest, David M. "Mutual Fund Performance Evaluation: A Comparison of Benchmarks and Benchmark Comparisons," *Journal of Finance*, **42** (1987), pp. 233–266.
119. Lynch, Anthony, and Musto, David. "How Investors Interpret Past Fund Returns," *Journal of Finance*, **58** (2003), pp. 2033–2058.
120. Mamaysky, Harry, Spiegel, Matthew, and Zhang, Hong. "The Dynamic Behavior of Mutual Fund Styles," unpublished manuscript, Yale University (2005).
121. Mamaysky, Harry, Spiegel, Matthew, and Zhang, Hong. "Improved Forecasting of Mutual Fund Alphas and Betas," *Review of Finance*, **11** (2007), pp. 359–400.
122. Massa, Massimo, Goetzman, William, and Rouwenhorst, Geert. "Behavioral Factors in Mutual Fund Flows," unpublished manuscript, Yale University (2001).
123. Merton, Robert C. "An Intertemporal Capital Asset Pricing Model," *Econometrics*, **41** (1973), pp. 867–887.
124. Miller, Tom W., and Gressis, Nicholas. "Nonstationarity and Evaluation of Mutual Fund Performance," *Journal of Financial and Quantitative Analysis*, **15** (1980), pp. 639–654.
125. Moskowitz, Tobias. "Discussion," *Journal of Finance*, **55** (2000), pp. 1695–1703.
126. Pastor, Lubos, and Stambaugh, Robert. "Investing in Equity Mutual Funds," *Journal of Financial Economics*, **63** (2000), pp. 351–380.
127. Pastor, Lubos, and Stambaugh, Robert. "Mutual Fund Performance and Seemingly Unrelated Assets," *Journal of Financial Economics*, **63** (2002), pp. 315–349.
128. Pollet, Joshua, and Wilson, Mungo. "How Does Size Affect Mutual Fund Behavior?" *Journal of Finance*, **63** (2008), pp. 2941–2969.
129. Poterba, James, and Shoven, J. "A New Investment Option for Taxable Investors," *American Economic Review*, **92** (2002), pp. 422–427.
130. Ross, Steven. A. "The Arbitrage Pricing Theory of Capital Asset Pricing," *Journal of Economic Theory*, **13** (1976), pp. 341–360.
131. Sharpe, William. "Mutual Fund Performance," *Journal of Business*, **39** (1966), pp. 119–138.
132. Sharpe, William. "Asset Allocation Management Style and Performance Measurement," *Journal of Portfolio Management*, **18** (1992), pp. 7–19.
133. Sharpe, William. "The Sharpe Ratio," *Journal of Portfolio Management*, **21** (1994), pp. 49–58.
134. Sirri, Eric, and Tufano, Peter. "Costly Search and Mutual Fund Flows," *Journal of Finance*, **53** (1998), pp. 1589–1622.
135. Stambaugh, Robert. "Analyzing Investments Whose Histories Differ in Length," *Journal of Financial Economics*, **45** (1997), pp. 285–331.
136. Treynor, Jack. "How to Rate Management of Investment Funds," *Harvard Business Review*, **43** (1965), pp. 63–75.
137. Treynor, Jack, and Mazuy, M. "Can Mutual Funds Outguess the Market?" *Harvard Business Review*, **44** (1966), pp. 131–136.
138. Treynor, Jack, and Black, Fisher. "How to Use Security Analysis to Improve Portfolio Selection," *Journal of Business*, **45** (1973), pp. 68–86.
139. Wenners, Russ. "A Matter of Style: The Causes and Consequences of Style Drift in Institutional Portfolios," working paper (2002).
140. Yan, Xuemin. "Liquidity, Investment Style and the Relation between Fund Size and Fund Performance," *Journal of Financial and Quantitative Analysis*, **43** (2008), pp. 741–768.
141. Zheng, Lu. "Is Money Smart? A Study of Mutual Fund Investors' Fund Selection Ability," *Journal of Finance*, **54** (1999), pp. 901–933.

27

Evaluation of Security Analysis

The selection of a portfolio of securities can be thought of as a multistage process. The first stage consists of studying the economic and social environment and the characteristics of individual companies to produce a set of forecasts of individual company variables. The second stage consists of turning these forecasts of fundamental data about the corporation and its environment into a set of forecasts of security prices and/or returns and risk measures. This stage is often called the *valuation process*. The third and last stage consists of forming portfolios of securities based on the forecast of security returns. Although, as we have seen in Chapter 26, a great deal of attention has been paid, both in the academic literature and in practice, to evaluating how well the entire process works, almost no attention has been paid to evaluating the components of the process. This is particularly surprising because the bulk of the evidence seems to indicate that the overall process does not work very well. The lack of extraordinary performance could be due to any of several causes, such as a lack of forecast ability, an inability to turn good forecasts of fundamental company data into good forecasts of returns, or a lack of ability to turn good forecasts of return into efficient portfolios. For example, it is perfectly possible that an organization has superior forecasting ability with respect to fundamental firm variables and market returns but does not capitalize on this information in forming portfolios.

In this chapter we are concerned with methods of analyzing how well an organization forecasts fundamental economic variables and how well it turns these forecasts into meaningful measures of security returns. To value a stock correctly, an organization must analyze and predict a large number of fundamental variables relating to each firm and the economy. In point of fact, the analysts at most institutions spend most of their time forecasting earnings (or growth in earnings). Because of this and because of the key role played by future earnings in any valuation scheme (see Chapter 18), we have selected forecasts of earnings per share as the fundamental firm variable examined in this chapter. The reader should keep in mind that the techniques we discuss for examining the accuracy of earnings estimates can be applied with a little imagination to forecasts of any fundamental variable. We start this chapter with a brief discussion of the sensitivity of price to earnings and an overall look at the accuracy of earnings estimates. Then we present techniques that should be useful in evaluating and diagnosing the errors in earnings forecasts. Finally, we study some techniques for examining the valuation process itself.

WHY THE EMPHASIS ON EARNINGS?

In Chapter 18 we saw that a firm's value was generally considered to be a function of dividends, growth, and risk. Forecasts of future dividends are usually prepared by applying a forecasted payout ratio to forecasted earnings. At least in the short run, payout ratios are easy to forecast and, to the extent they vary from historical levels, they usually do so as a function of earnings changes. We have already devoted a large amount of material to the analysis and forecasting of risk (Chapters 7 and 8). Thus the key remaining variable is the forecast of future earnings.

In Chapter 19 we showed that an ability to forecast future earnings can allow an excess return to be earned, even in the absence of a complex valuation model. For example, we saw that the 30% of firms that had the largest increase in earnings offered the investor a risk-adjusted excess return of 7.48% over a 13-month period, while those in the lower 30% offered a risk-adjusted excess return of −4.93%.[1]

The ability to earn an excess return by correctly forecasting earnings implies that the market's forecast of earnings (which determines price) is not perfectly accurate. Further evidence of this is supplied by the fact that the excess return we can earn from a perfect forecast of earnings becomes smaller and smaller as the end of the fiscal year approaches.[2] This is consistent with the market's estimate of future earnings becoming more and more accurate as information is released during the fiscal year.

A very good proxy for market expectation of future earnings is the consensus forecasts of security analysts. If the average forecast of security analysts is close to market expectations, then one should not be able to purchase stock on the basis of these expectations (e.g., forecasted growth) and make an excess return. Most empirical evidence strongly suggests that this is, in fact, true.[3]

On the other hand, if the consensus forecasts are a good proxy for market expectations and one can forecast with more accuracy than the average analyst, then one should be able to make an excess return. In Chapter 19 we saw that this was, true. The next logical question to ask is how large has the error in consensus forecasts been. The answer is, quite large. If we examine forecasts made nine months before the end of the fiscal year, we find that for the 30% of the companies for which analysts most overestimated growth in earnings, their average error in forecasting growth was 63.6%. If we examine the 30% of the companies for which analysts most underestimated growth, we find their estimates were off, on average, by −38.9%. As the end of the fiscal year approaches, these errors shrink. But three months before the end of the fiscal year, they were still quite large: +26.4% and −27.0%, respectively.

It would be interesting to see how well individual analysts have performed compared to the consensus estimates. Unfortunately, no such studies exist. However, there are three studies of how accurately individual analysts forecast compared to simple extrapolations of past earnings.

Cragg and Malkiel (1968) analyzed predictions of long-term (five-year) growth rates prepared on each of a large sample of firms by analysts at five leading institutions. They concluded that there seems to be no clear-cut ability of the institutions examined to outperform simple extrapolations of historical growth rates.

---

[1]See Elton, Gruber, and Gultekin (1981) and Francis and Schipper (1999) for evidence that knowledge of future earnings can lead to excess returns.

[2]See Elton, Gruber, and Gultekin (1981) for additional evidence.

[3]Elton, Gruber, and Gultekin (1981) and Fried and Givoly (1982) provide evidence to support these statements. On the other hand, LaPorta (1996) and Dechow and Sloan (1997) provide evidence that while consensus forecasts are incorporated in market prices, errors in long-term consensus forecasts can be identified and allow excess returns in the short run.

Elton and Gruber (1972) analyzed the ability of the analysts at three financial institutions to predict earnings nine months before the end of the fiscal year. They found that, of the three institutions examined, one performed slightly worse than historical extrapolation of past earnings, two performed slightly better, but none of the differences was statistically significant at even the 10% level. Brown and Rozeff (1978) investigated the performance of Value Line estimates of earnings, once again comparing them with historical extrapolation methods. They found that Value Line outperformed the extrapolation techniques at a statistically significant level. A reasonable conclusion to draw from this evidence is that, while it is not easy to outperform historical extrapolation, there are individuals and perhaps institutions that might be able to do so. It may also be true that there are individuals and institutions that can outperform the consensus forecasts.

It is surprising, in light of the impact of the accuracy of earnings forecasts on stock selection and in light of the tremendous resources that financial institutions devote to the preparation of earnings forecasts, that more resources have not been devoted to the evaluation of earnings estimates. This is the subject with which the next section of this chapter deals.

THE EVALUATION OF EARNINGS FORECASTS

Although very little has been written about the evaluation of the estimates of security analysts, there is a broad literature in economics on the evaluation of forecasts. We draw heavily on this literature and, in particular, on the work of Henri Thiel in this section. We start by examining a meaningful overall measure of the accuracy of earnings forecasts. Then we look at both graphical and numerical techniques for diagnosing the sources of forecast error. We will end this section with an argument that any evaluation of earnings forecasts should be performed relative to the consensus (average) forecast of earnings. Anticipating this discussion and to provide a benchmark against which to measure analysts' performance, we provide data on the performance of the consensus forecast for several error measures discussed in this section.[4]

Overall Forecast Accuracy

To evaluate earnings forecasts in an exact manner, one should really have a loss function that measures the loss caused by any size error in the forecast. While this is indeed desirable, in many fields the evaluation of estimates must be performed in the absence of an explicit loss function. The most frequently assumed loss function is the quadratic, and the most frequently assumed measure of loss is the mean squared error.[5]

The mean squared forecast error for any set of forecasts can be easily computed as

$$\text{MSFE} = \frac{1}{N}\sum_{i=1}^{N}\left(F_i - A_i\right)^2$$

where

MSFE is the mean squared forecast error

[4]The data used for the results reported in this chapter were extracted from the I/B/E/S database. We include the consensus forecasts for all corporations with a December fiscal year, which were followed by three or more analysts for a three-year period. Most tables are based on a total of 1,242 consensus forecasts.

[5]See Thiel (1964, 1966).

F_i is a forecast of the earnings per share for firm i

A_i is the actual earnings per share that occurs for firm i

The mean squared forecast error is often developed in terms of the change in earnings. Define

P_i as the predicted change in earnings

R_i as the realized change in earnings

H_i as the level of earnings at the time the forecast was made.

Then,

$$P_i = F_i - H_i$$

$$R_i = A_i - H_i$$

The mean squared error in terms of the change in earnings can be written as

$$\text{MSFE} = \frac{1}{N}\sum_{i=1}^{N}\left(P_i - R_i\right)^2$$

The same MSFE results whether we perform the analysis in terms of the predicted change in earnings or the predicted level of earnings.[6]

It will prove convenient for error diagnosis to examine the forecast error as the error in forecasting the change in earnings. The mean squared forecast error can be used to rank forecasting techniques. However, it would be useful to scale the MSFE so that its value has a natural interpretation. One useful way to scale it has been suggested by Thiel (1964, 1966). This measure, often referred to as Thiel's inequality coefficient (TIC), involves dividing the MSFE by the sum of the squared change in earnings, or

$$\text{TIC} = \left(\sum_{i=1}^{N}\left(P_i - R_i\right)^2\right) \Bigg/ \left(\sum_{i=1}^{N} R_i^2\right)$$

Notice that two values of this measure have an easily interpreted economic meaning. If the predicted change in earnings always equaled the realized change in earnings (perfect forecasting), then the numerator would be 0 and TIC would be 0. Thus a value of TIC equal to 0 implies perfect forecasting ability. If the predicted change in earnings always equaled 0, then P_i would equal 0 and TIC would equal 1. A value of TIC equal to 1 implies that the forecasts are exactly as accurate as a forecast of no change in next period's earnings. TIC allows us to obtain a sense of how well a forecaster performs, even before comparisons are made with other forecasters. A value below 1 indicates that the forecaster outperforms the naive no-change model. A value larger than 1 indicates that the forecasts could not outperform the most naive of all forecasting models.

6

$$\text{MSFE} = \frac{1}{N}\sum_{i=1}^{N}\left(P_i - R_i\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left[F_i - H_i - \left(A_i - H_i\right)\right]^2 = \frac{1}{N}\sum_{i=1}^{N}\left(F_i - A_i\right)^2$$

Examination of the last term makes it clear that the MSFE of the change in earnings is identical to MSFE in terms of level of earnings.

Table 27.1 TIC over Time

Months before final forecast	11	10	9	8	7	6	5	4	3	2	1	0
TIC	0.75	0.70	0.62	0.54	0.49	0.44	0.41	0.35	0.28	0.26	0.20	0.15

Table 27.1 presents the values of TIC for the consensus estimates over our three-year period. Notice that 11 months before the last consensus forecast, the average analyst has a TIC value of 0.75, indicating that the performance of the consensus forecast is slightly better than the naive forecast. By the time analysts prepare their last forecast, the value of TIC has fallen to 0.15, indicating a great deal of accuracy. Table 27.1 also shows an extremely regular decrease of TIC for successive forecasts. The coefficient of determination between TIC and time is 0.99. As Elton, Gruber, and Gultekin (1981) show, this same pattern occurs in each year as well as on average over the entire period.

Before leaving this section, we should mention an alternative to examining forecasts in terms of the error in forecasting earnings. Some researchers have suggested that forecasts be examined in terms of the *percentage* error in forecasting earnings. Ultimately, the test of which is correct depends on whether losses or gains are a function of the amount or the percentage by which earnings are misestimated. While this decision is up to the user, he or she should be aware that if percentage errors are used, results tend to be dominated by the huge percentage errors usually found in companies with very small earnings. Very small size earnings (a small denominator in the percentage calculation) make misestimation of any size appear quite serious. A problem also arises when the company has zero or negative earnings.[7] The analysis in this section can easily be recast in terms of percentage earnings error. The reader must decide which is the most relevant criterion.

Diagnosis of Forecasting Errors

There are infinitely many ways to examine forecast errors to learn more about them and perhaps correct future forecasts for their deficiencies. In this section we first present a diagrammatic scheme that can be used to learn a great deal about the pattern of forecast errors, and then we present some numerical techniques for computing diagnostics.

Graphical Analysis One of the simplest, and yet most revealing, techniques for examining the pattern in forecast errors is the Prediction Realization Diagram (PRD) proposed by Thiel (1964, 1966). This diagram is simply the plot of the predicted change in earnings against the realized change. The predicted change is plotted along the line that lies at a 45° angle to the horizontal axis, and actual change is plotted along a line that lies at a $-45°$ angle to the horizontal axis. This is shown in Figure 27.1.

If we plot in this space the forecasted change in earnings versus the realized change, we can learn quite a lot about the type of forecast errors being made. Notice that if a point lies on the horizontal straight line, it indicates that the forecast change was exactly equal to the actual change. To the extent that points lie above the horizontal straight line, it indicates that estimates were too high. To the extent that points lie below the horizontal line, it indicates that estimates were too low. Now let us take a closer look at what each section of the graph represents. A point lying in section I of the PRD indicates that the forecaster successfully

[7]When earnings are zero, the percentage change in earnings will be infinite. When earnings are negative, the meaning of percentage changes in earnings is ambiguous.

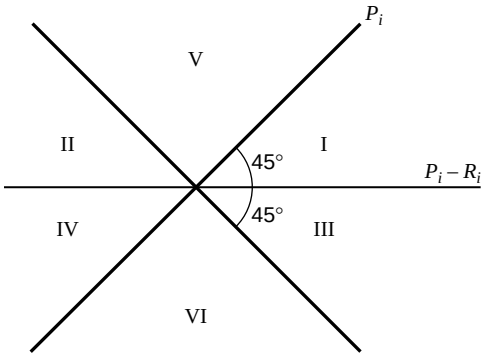


Figure 27.1 Prediction Realization Diagram.

predicted that earnings would increase but that the size of the increase was overestimated. A point lying in section II indicates that the analyst successfully predicted a decrease in earnings but that the size of the decrease was underestimated (earnings were overestimated). If a point lies in section V, it indicates that the analyst predicted the wrong direction for the change in earnings. That is, the analyst predicted they would increase when they, in fact, decreased. Sections III, IV, and VI are analogous to sections I, II, and V. Section III represents a successful prediction of an increase in earnings but an underestimate of the size of the increase. Section IV represents a successful prediction of a decrease but an overestimate of the size of the decrease. Finally, a point in section VI represents a forecast of a decrease in earnings when they actually increased. Sections V and VI indicate that the analyst misestimated the direction of the change in earnings movements, whereas the other sections indicate the analyst got the direction right but the size wrong.

Examination of a group of forecasts on the PRD can reveal quite a lot of information about the source of error in analysts' forecasts. In Figures 27.2 and 27.3 we have constructed two hypothetical patterns that might be observed. Figure 27.2 presents the case where a forecaster is consistently optimistic. The forecaster consistently overestimates earnings changes when they are positive (section I) and consistently underestimates the size of a negative change (section II) or actually predicts a positive change when changes are negative (section V).

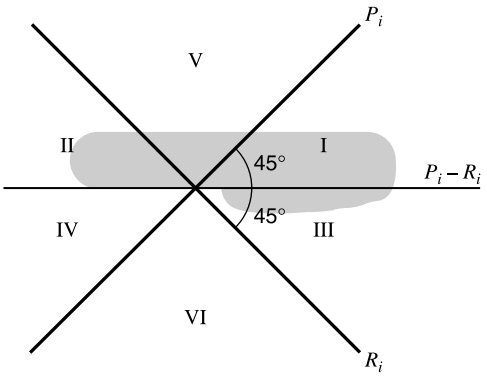


Figure 27.2 Prediction Realization Diagram: Optimistic forecaster.

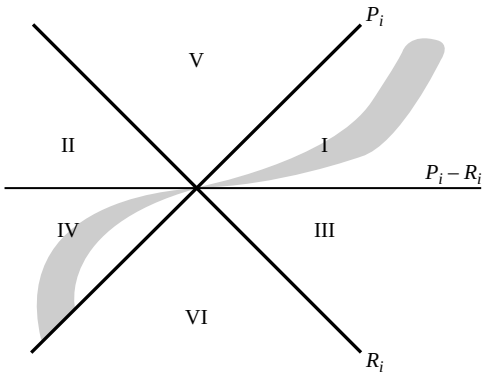


Figure 27.3 Prediction Realization Diagram: The overreactor.

A more interesting pattern is revealed in Figure 27.3. In this diagram we find the profile of an analyst who is excellent at predicting the direction of change. However, the analyst overreacts to change by becoming overoptimistic as large positive changes are expected and overpessimistic as large negative changes are expected. This can be seen by the fact that estimates lie further and further from the horizontal axis as the actual level of earnings change increases.

These are only two of the many potential patterns of forecaster behavior that can be seen through the PRD. The reader is encouraged to construct other series of points on this diagram and to interpret their meaning.

Numerical Analysis While the graphical analysis of forecast errors is extremely useful, there are also several analytical decompositions of the mean squared forecast error that can provide useful insight into the sources of forecasting error. Let us take, as an example, a firm that has a collection of analysts making forecasts of a broad group of stocks in the economy. We discuss two meaningful decompositions of these errors. The first is based on the level of aggregation at which errors occur, while the second looks at the forecast error in terms of the characteristics of the forecasters.

Error Decomposed by Level of Aggregation It would be extremely useful to determine at what level of aggregation errors in the forecasts occurred. One scheme for analyzing the level separates earnings errors into three components;

$$\begin{aligned}\text{MSFE} &= \frac{1}{N}\sum_{i=1}^{N}\left(P_i - R_i\right)^2 \\ &= \left(\overline{P} - \overline{R}\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left[\left(\overline{P}_a - \overline{P}\right) - \left(\overline{R}_a - \overline{R}\right)\right]^2 \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}\left[\left(P_i - \overline{P}_a\right) - \left(R_i - \overline{R}_a\right)\right]^2\end{aligned}$$

where

$\overline{P}$ = mean value of P_i across all stocks followed by all analysts

$\overline{R}$ = mean value of R_i across all stocks followed by all analysts

$\bar{P}_a$ = mean value of P_i for industry a to which i belongs; each industry will have a different value of $\bar{P}_a$

$\bar{R}_a$ = mean value of R_i for each industry in turn

The first term measures how much of the forecast error is due to inability of the analysts, in total, to predict what average earning will be for the economy. This term is simply the squared difference between the average predicted change in earnings and the average realized change in earnings. The second term is a measure of how much of the total error is due to the individual analysts misestimating the differential performance of particular industries from the average for the economy.

Let us examine this term in more detail. For each firm (i), the difference between the mean predicted change in earnings for the industry to which it belongs ($\bar{P}_a$) and the mean predicted change in earnings for all firms ($\bar{P}$) is calculated. The same term is calculated for actual change in earnings ($\bar{R}_a - \bar{R}$). The difference between the two is squared and summed for all firms. Then the average value of this term is taken. The third term measures how much of the error is due to analysts not being able to predict the difference in performance of the individual stocks they follow from the appropriate industry average. The first part of this term is the difference between the predicted change in earnings for an individual stock P_i and the average predicted change for a stock in the industry to which i belongs. The second part has the same meaning but deals with realizations. These differences are squared, summed, and then averaged.

By dividing through both sides of the equation by the MSFE, we express each source of error as a fraction of the total mean squared forecast error. Diagnosing the source of the error can be of great significance to the firm. For example, if the major source of error arises from misestimated aggregate earnings, then the company should concentrate more effort on preparing its forecasts of the general economy. If the analysts are provided with better information about the aggregate level of earnings and are explicitly encouraged to use this information, then improvement should occur in the firm's forecasting effort. Assuming that each analyst follows one industry or a group of closely related industries (economic sector), then a large value for the second term points to an error in understanding the economics of alternative industries. Large values of the third term indicate that errors are associated with being unable to differentiate between the performance of individual companies even when mistakes in forecasts of the level of the economy and industries are removed.

Obviously, this same type of decomposition can be repeated for all individual analysts with their error decomposed into their misestimate of how the stocks they follow will do, on average, and their inability to differentiate the performance of the companies they follow.

In Table 27.2 we present the decomposition of the mean square error by level of aggregation for the set of consensus forecasts discussed earlier. Perhaps the most striking aspect of this table is the small percentage of error that is due to misestimating the performance of the economy (the average company). This source of error never exceeds 3% of the total mean squared error. The percentage of error due to industry misestimates starts at 36.5% in January and declines continuously to 17.6% by the end of the fiscal year. Consequently, the error due to misestimating individual companies grows from 61.8% to 80.5% over the year. We have already seen that the size of analysts' errors shrinks over the year. Now we see that while analysts become more accurate in forecasting both industry and company errors, their ability to forecast industry influences grows relative to their ability to forecast company performance over the year.

Errors Decomposed by Forecast Characteristics There is a second type of decomposition of forecast errors that is meaningful to management. This decomposition looks for

Table 27.2 Percentage Error in Earnings Change by Level of Aggregation

	Economy	Industry	Company
January	1.7	36.5	61.8
February	1.8	36.4	61.8
March	1.9	35.6	62.4
April	1.9	33.7	64.4
May	2.4	33.4	64.2
June	2.7	31.8	65.6
July	2.8	31.7	65.5
August	2.8	30.9	66.2
September	3.0	28.2	68.9
October	2.7	27.2	70.1
November	2.2	23.6	74.3
December	1.9	17.6	80.5

the pattern of mistakes and is a numeric analogue to the graphical analysis presented earlier. We can write this decomposition as

$$\text{MSFE} = \frac{1}{N}\sum_{i=1}^{N}\left(P_i - R_i\right)^2$$

$$\text{MSFE} = \left(\overline{P} - \overline{R}\right)^2 + \left(1-\beta\right)^2\sigma_p{}^2 + \left(1-\rho\right)^2\sigma_R{}^2$$

where

- β is the slope coefficient of the regression of R on P
- ρ is the correlation of P and R
- σ_P^2 is the variance of P
- σ_R^2 is the variance of R

The first term in this equation ($\overline{P} - \overline{R}$) represents bias. This is the tendency of the average forecast to either overestimate or underestimate the true average. The second term represents inefficiency or the tendency for forecasters to be systematically overoptimistic (or insufficiently optimistic) about good (or bad) events. If the beta of actual earnings on forecasted earnings is greater than 1, then forecasts are overestimates of earnings at high values and underestimates at low values of actual earnings. If beta is less than 1, then analysts underestimate earnings when they are high and overestimate earnings when they are low. The final component of this equation is the random disturbance term.

When we apply this decomposition to consensus estimates, we find some interesting results. The vast majority of analysts' errors are random rather than systematic in nature. In all months over 91%, and in one-half of the months over 94%, of the MSFE arise from random error. Inefficiency as well as bias contribute very little to MSFE. Furthermore, betas are close to 1 and vary randomly around 1. There does not seem to be a systematic tendency of analysts to get overexcited or overly cautious about potentially good performance on the part of firms.

The Evaluation of Earnings Forecasts—Again

While it is important to examine the accuracy of earnings estimates and to diagnose where errors are being made, another step should be taken. Forecast errors can more meaningfully be judged relative to some benchmark than they can on an absolute basis. The need

for a benchmark is easy to see. It is less difficult to forecast earnings for some stocks and for some industries than it is for others. For example, the earnings of public utilities are more stable and easier to forecast than are the earnings for electronics manufacturers. If a benchmark is not used, the forecaster who follows utilities or the firm that specializes in utility stocks will be judged a better forecaster (if its ability is about the same) than the forecaster who follows electronics stocks.

Thus one quality we would like a benchmark to have is to adjust for the difficulty of the forecasting process. A second quality we would like a benchmark to have is to represent an absolute base such that forecasting ability above the benchmark can be potentially transformed into superior security selection, while performance below the benchmark is unlikely to lead to superior security selection. Fortunately, a benchmark exists that satisfies both these criteria. It is the consensus forecast introduced in Chapter 19. The accuracy of the consensus forecast reflects how easy or difficult it is to forecast the earnings for a particular company or group of companies. In addition, as we have already seen, the price of any stock reflects (incorporates) the consensus forecast. The ability to forecast with no more accuracy than the consensus should not lead to a superior return, while the ability to forecast with more accuracy should.

Thus we propose that the benchmark against which all analysts and forecasts be judged be the consensus forecast. As a first step in analyzing the performance of any analyst or group of analysts (institution), their mean squared forecast error can be computed and compared directly with the mean squared forecast error for the consensus forecast. The consensus mean squared forecast error should be computed over that same set of corporations for which the analyst or institution prepared forecasts. Then each of the diagnostics discussed earlier in this chapter can be computed for the consensus forecasts and compared directly with the diagnostics of the institution's forecasts. Alternatively, consensus forecasts can be used as the benchmark forecast and individual forecasts compared against it. This would involve defining realizations as the consensus forecast of change in earnings.

EVALUATING THE VALUATION PROCESS

The valuation process converts a set of forecasts about company fundamentals and economic data into a set of forecasts of market variables or a recommended course of action to take with respect to individual securities. The number of books and articles written about the valuation of securities far exceeds the number of publications on modern portfolio theory. It is surprising that, given the large number of publications on how to value securities, almost nothing has been written about how to value the valuation process itself.

Before turning to an evaluation of the valuation process, let us spend a little time thinking about what form the output from the valuation process should take. The problem becomes rather simple when we realize that the output from the valuation process is the input to portfolio analysis. We know what we need to perform portfolio analysis. We need estimates of the expected return for each stock, the variance of the return on each stock, and the correlation of returns between each pair of stocks. Chapters 7 and 8 dealt with alternative models for predicting correlations between securities as well as techniques for evaluating these alternative models. We mentioned at that time that we felt it unlikely that the analyst would ever produce direct estimates of correlation coefficients.

What analysts can produce are estimates of expected returns, perhaps variances, or estimates of the parameters of at least one of the models from Chapters 7 and 8 that can be used to estimate variance and covariances. For illustrative purposes, let us assume that analysts are preparing estimates of expected returns and betas. Their estimates of beta may well involve subjective modification of historical or fundamental betas.

The question then remains, given that analysts produce estimates of the relevant risk and return parameters, how do we evaluate the quality of these estimates?[8]

Evaluating the Valuation Process with a Full Set of Outputs

There are really three steps that can and should be taken in evaluating the valuation process:

1. How well does each output from the valuation process predict the future?
2. If the output is examined in a simple way, does the output lead to undervalued and overvalued stocks being correctly identified?
3. If the output is used in an optimal manner, does it produce good results?

Let us now examine each in turn. The first step is to see if there is any predictive content in the output for the valuation process. For example, one type of output should be the expected return for each security. One question to analyze is, how well does expected return forecast future returns, and what are the sources of error in the forecast? We have already designed the system for analyzing this question in the first part of this chapter. For all of the evaluation procedures outlined in this part of the chapter, we can simply substitute expected rate of return for change in earnings. For the naive model against which to judge expected rate of return, we can substitute the historical average rate of return on the stock for the consensus forecast of earnings.

Any other individual output from the valuation process can be analyzed in an analogous manner. For example, beta estimates could be used in any of the diagnostic procedures outlined in the first part of this chapter.

Let us assume that there is some predictive content in one or more of the outputs from the valuation process. Where do we go from there? The next step is to begin to analyze the output in combination. The simplest way to do this is to examine the output in expected return beta space and to see how stocks that appear to be underpriced (or overpriced) perform in subsequent periods. To be more specific, assume that analysts as of December 31, 2011, have forecasted the expected return and beta for a group of stocks for the year 2012. Based on expectations about returns and betas for the year 2012, an expected security market line could be constructed. The distance that a particular security lies above this line is a measure of its attractiveness as a candidate for purchase. The distance above or below the expected security market line is usually referred to as a stock expected alpha. This is a measure of how much more or less than its equilibrium return a stock is expected to earn. At the end of 2012, the actual return and beta for each stock can be measured as well as the security market line for 2012. This allows the computation of an actual alpha for each stock for 2012. The expected alpha for each stock can now be compared with the alpha that occurred for each stock. In particular, the predictive power of expected alpha can be compared with the naive prediction of all future alphas equal to zero. If the expected alphas predict better than this model, there is informational content in the valuation process.

The final step in evaluating the output from the valuation process is to see if employing this output with the portfolio optimization rules of Chapters 6 and 9 leads to the selection

[8]In this section we are presenting techniques for the evaluation of the valuation process given the quality of inputs to the valuation process. This should not be disturbing, because the previous section dealt with evaluating these inputs.

of portfolios that perform well. Using the appropriate techniques from Chapters 7 and 9, we select that portfolio that would be optimum if the input from the valuation process were correct. We can then use any of the techniques outlined in the previous chapter to evaluate the portfolio we have selected. It is hoped that this portfolio will outperform naive strategies such as buying an index fund with the same risk.

Although we could end the discussion of evaluation of the valuation process at this point, one more step should be taken. Using the techniques of the previous chapter, the portfolio discussed earlier should be compared with the portfolio actually held by the institution performing this analysis. If the institution's portfolio does not perform as well, it indicates that the portfolio manager is not making optimum use of the information supplied to him. If the manager's portfolio performs better, it indicates that he or she is introducing additional information not contained in the output of the valuation process.

Evaluating the Output of the Valuation Process: Incomplete Information

To perform portfolio analysis properly, one needs a full range of inputs on expected return, variances, and covariances. While more and more firms are recognizing this and encouraging their analysts to provide data in this form, the majority of firms have a much simpler form for output from the valuation process.

Many firms have their analysts provide data to portfolio managers (or provide data to the firm's clients) in terms of a recommendation to either buy, sell, or hold particular stocks.[9] This is less satisfactory than output provided in terms of rate of return forecasts. It forces the analyst to compress a continuous rating of securities into a three-point scale. This prevents the analyst from passing along information to the portfolio manager or, perhaps worse, gives the analyst an excuse for not developing the information. The best evaluation that can be done in a case like this is to examine the performance of each of the three groups of stocks to see if the groupings contain information.

As an example of the type of analysis that might be done, let us assume that we have a group of stocks for which buy-sell-hold recommendations have been made as of December 31, 2011. The firm has a one-year time horizon, and it is now December 31, 2012. Then the returns on each stock for the year 2012 can be plotted in return beta space. Furthermore, the best estimate of the security market line that existed in 2012 can be plotted on the same diagram. The alphas or distances above and below the line could be computed for each stock.[10] Figure 27.4 illustrates this analysis.

It is hoped that the majority of the buy recommendations lie above the line and the majority of the sell recommendations lie below the line. The difference between the rate of return on the stock and the return expected on each stock given the security market line can be calculated for each buy recommendation as well as for each hold and sell recommendation.[11] The average for each group can be computed. It should be positive for the buy recommendations, close to zero for the hold, and negative for the

[9]Some firms use a five-point scale rather than a buy-sell-hold recommendation, but the evaluation will be similar to that described in this section.

[10]The estimate of the security market line should be made from all stocks, not just from those stocks followed by the firm.

[11]This distance is simply the vertical distance between the line and the point.

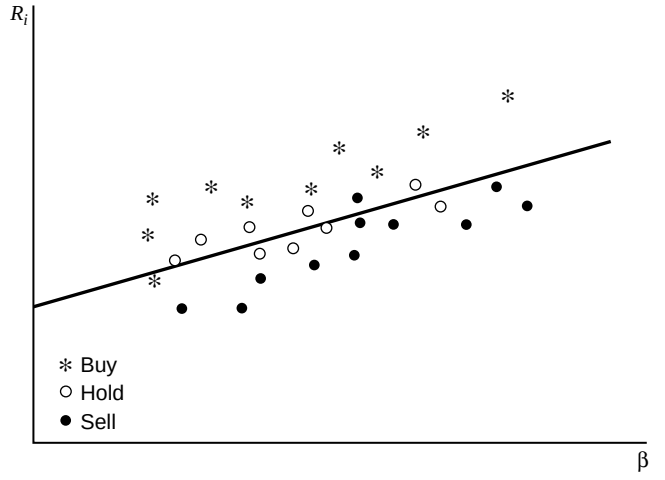


Figure 27.4 Examination of buy-hold-sell recommendations.

sells. Notice that, even in computing the average distances for each group, we have implicitly made an assumption. The assumption is that in putting together a portfolio, the portfolio manager will place an equal dollar amount in each buy and an equal dollar amount in each sell. This is not an optimal course of action, even if the buy and sell ratings are perfectly accurate. There are bound to be differences in the relative rankings of stocks within each category, and this procedure fails to take account of these differences. However, the naive procedure of assuming equal investment can be used to get an indication of whether there is information in the estimates of buy, hold, and sell.[12]

CONCLUSION

Almost all the emphasis in the evaluation process has been placed on evaluating the performance of portfolios held by financial institutions. In light of the fact that the performance of these portfolios has been unsatisfactory, it is important to start examining the steps in the portfolio management process to see if there is information that is not being used. In this chapter we have suggested a series of steps for doing so. The first step is to examine the forecasts of fundamental data on corporations to see if they contain information. The second step is to evaluate the output from the valuation process to see if it has taken advantage of any information contained in the security analyst's basic forecast. The final step is to compare portfolios selected in an optimal fashion from the output of the valuation process with portfolios selected by portfolio managers to analyze what the portfolio manager adds to the process. It is only by breaking the portfolio selection and management process into stages that a firm can find what it does well and what it does poorly in the hope of improving portfolio performance.

[12]This naive procedure is necessary if one wishes to isolate the effect of information in the buy, hold, or sell recommendations. If one wishes to simultaneously evaluate risk variables, a more complex procedure is possible.

QUESTIONS AND PROBLEMS

1. Assume that a brokerage firm concentrates on a few closely related industries. It has produced a set of estimates of earnings for 1985 and subsequently recorded the earnings that actually occurred. These data are given below:

Industry	Firm	Previous Earnings	Estimated Earnings	Actual Earnings
A	1	1.05	1.10	1.05
	2	1.32	1.37	1.35
	3	3.50	4.25	3.25
B	4	2.06	2.10	2.12
	5	2.08	2.13	2.12
	6	2.60	3.25	2.80
	7	1.07	1.06	1.06
C	8	2.00	2.70	2.40
	9	0.55	0.52	0.54
	10	1.18	1.16	1.20

A. Plot these points on a Predictive Realization Diagram. What can we learn about the forecast pattern of this firm from the PRD?

B. Calculate the mean square forecasted error for this firm.

C. Decompose the error by level of aggregation. That is, determine what percentage of the error was due to the inability to forecast earnings for this sector of the economy, what percentage was due to an inability to forecast each industry, and what percentage was due to an inability to forecast differences for each firm.

D. Examine another level of decomposition. Assume that there are three analysts, each following one industry. What is the mean squared error of each analyst? How much of the error of each analyst is due to the analyst's inability to predict the future of the industry followed, and how much is due to an inability to differentiate between the firms in the industry?

E. Decompose the error by forecast characteristics. Find what percentage of the error is due to bias, what percentage is due to variance, and what percentage is due to covariance.

BIBLIOGRAPHY

1. Brown, Lawrence, and Rozeff, Michael. "The Superiority of Analysts' Forecast as Measures of Expectations: Evidence from Earnings," *Journal of Finance*, **XXXIII,** No. 1 (March 1978), pp. 1–16.
2. Brown, Stephen J., and Warner, Jerold B. "Measuring Security Price Performance," *Journal of Financial Economics*, **8,** No. 3 (Sept. 1980), pp. 205–258.
3. Cragg, J. G., and Malkiel, B. G. "The Consensus and Accuracy of Some Predictions of Growth of Corporate Earnings," *Journal of Finance*, **23,** No. 1 (March 1968), pp. 67–84.
4. Dechow, P., and Sloan, R. "The Relation between Analysts' Forecasts of Long-Term Earnings Growth and Stock Price Performance Following Equity Offerings," *Contemporary Accounting Research*, **17,** No. 1 (2000), pp. 1–32.

5. Dybvig, Philip H., and Ross, Stephen A. "The Analytics of Performance Measurement Using a Security Market Line," *Journal of Finance*, **40,** No. 2 (June 1985), pp. 401–416.
6. ——. "Performance Measurement Using Differential Information and a Security Market Line," *Journal of Finance*, **40,** No. 2 (June 1985), pp. 383–400.
7. Elton, Edwin J., and Gruber, Martin J. "Earnings Estimates and the Accuracy of Expectational Data," *Management Science*, **18,** No. 2 (April 1972), pp. 409–424.
8. Elton, Edwin J., Gruber, Martin J., and Gultekin, Mustafa. "Expectations and Share Prices," *Management Science*, **27,** No. 9 (Sept. 1981).
9. ——. "Professional Expectations: Accuracy and Diagnosis of Errors," working paper, New York University (1983), pp. 975–987.
10. Elton, Edwin J., Gruber, Martin J., and Grossman, Seth. "Discrete Expectational Data and Portfolio Performance/Comment," *Journal of Finance*, **41,** No. 3 (July 1986), pp. 699–714.
11. Fama, Eugene F., and French, Kenneth R. "Size and Book-to-Market Factors in Earnings and Returns," *Journal of Finance*, **50,** No. 1 (March 1995), pp. 131–155.
12. Francis, Jennifer, and Schipper, Catherine. "Have Financial Statements Lost Their Relevance?" *Journal of Accounting Research*, **37,** No. 2 (1999), pp. 319–352.
13. Fried, Dov, and Givoly Dan. "Financial Analysis" Forecasts of Earnings: A Better Surrogate for Market Expectations," *Journal of Accounting and Economics*, **4** (1982), pp. 85–107.
14. LaPorta, Rafael. "Expectations and the Cross-Section of Stock Returns." *Journal of Finance*, **51** (1996), pp. 1715–1742
15. Thiel, Henri. *Optimal Decision Rules for Government and Business* (Amsterdam: North-Holland, 1964).
16. ——. *Applied Economic Forecasting* (Amsterdam: North-Holland, 1966).

28

Portfolio Management Revisited

Throughout this book we have presented analyses and models that have major implications for the way money should be managed. Some of the analyses involved forecasts of economic or market characteristics for securities and optimal techniques for portfolio construction based on these forecasts. Other chapters imply that securities prices are in equilibrium or almost in equilibrium, and the investor should hold some sort of aggregate portfolio. We have not resolved this issue, nor has the investment community resolved it.

It seems worthwhile to take a last look at how money management has evolved and how the financial community has dealt with these issues. In this chapter we will discuss the major investment strategies that modern portfolio managers follow. For each strategy, we will discuss the assumptions under which the strategy should be successful. In addition to supplying a review of current approaches to investment analysis, this chapter should help the reader integrate earlier chapters of this book. Let's start by briefly reviewing the basic approaches to investment management. In general terms the approaches are labeled as passive management and active management. We present more of their overall characteristics in Figure 28.1. In earlier editions of this book we briefly described each approach. In recent years, however, because the investment process has evolved and resulted in the structuring of so many new investment management styles and products, we are going to devote more discussion to their analysis.

The chapter is divided into four sections. In the first section we discuss management styles for stock portfolios. In the second we discuss management styles for bond portfolios. The distinction is made for two reasons:

1. Types of investment products have developed in the bond area that have not yet been developed for stocks.
2. Because of both finite lives and liquidity considerations, bonds present some special challenges.

In the third section of this chapter we introduce some of the concepts of managing a portfolio when the manager is concerned with meeting a set of liabilities. In the final section we discuss the bond–stock mix.

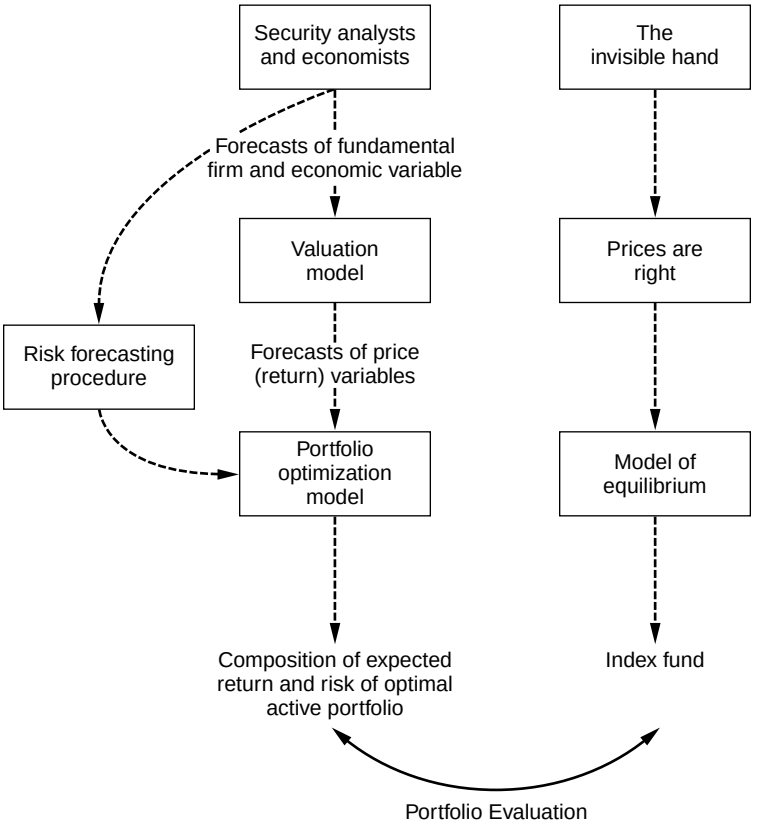


Figure 28.1 Modern version of traditional approach.

MANAGING STOCK PORTFOLIOS

In "olden days" (5 or 10 years ago), it was easy to discuss management styles for common stocks. One was a passive manager or an active manager. The passive manager held a market index (e.g., the Standard and Poor's (S&P) 500), and the active manager did something else. Although it would be easy to discuss management styles in this way, we would be overlooking recent developments in the area of passive management. Instead, we distinguish active versus passive management on the basis of whether action is predicated on forecast data. As we will see, even with this definition, the line between passive and active management becomes increasingly fuzzy.

Passive Management

Funds under passive management have grown rapidly and reached significant size. By 2011, assets in passive portfolios comprised approximately 13% of the total assets managed by investment companies. The simplest case of passive management is the index fund that is designed to replicate *exactly* a well-defined index of common stock, such as the S&P 500. The fund buys each stock in the index in exactly the proportion it represents of the index. If IBM constitutes 4% of the index, the fund places 4% of its money in IBM stock.

The standard (Sharpe–Lintner–Mossin) capital asset pricing model could be considered the theoretical justification for such a fund, if one were willing to accept the S&P 500 as a suitable proxy for the market portfolio of risky assets. Perhaps a more practical justification is that index funds have outperformed more than 50% of active managers. One of the major companies evaluating manager performance estimated in 2010 that during the past 20 years the S&P index outperformed more than 70% of active managers.

Although exact replication is the simplest technique for constructing an index fund, many index funds are not constructed this way. Managers of index funds must face a series of decisions in designing a fund. These decisions involve the trade-off between accuracy in duplicating the index (called tracking error) and transaction costs. Does the manager buy all 500 stocks in market proportions, or are some of the stocks with the smallest market weight excluded to save on transaction costs? How are periodic dividends reinvested to balance savings on transaction costs versus imperfections in tracking the index? How much cash should be kept on hand to accommodate withdrawals or as a result of cash inflows?[1] The more cash, the lower the transaction costs, but the less perfectly the index is tracked.

There are three commonly used approaches in constructing an index fund. Each makes a different trade-off between accuracy in duplicating the index and transaction costs. These three approaches can be summarized as follows:

1. Hold each stock in the proportion it represents of the index.
2. Mathematically form a portfolio of not more than a specified number of stocks (e.g., 300), which best tracks the index historically. Standard mathematical programming algorithms can be used to do this.
3. Find a smaller set of stocks that matches the index in the percentage invested in a prespecified set of characteristics (e.g., same percentage in industrial, utility, and financial stocks). Some of the frequently used characteristics are sector, industry, quality, and size of capitalization.

Although some index funds replicate market weights exactly, many use a combination of the first approach and either the second or third approach. Market weight matching is most likely to be used by funds that match an index of large capitalization stocks, such as the S&P 500 index. For funds that match a much broader index (e.g., the Wilshire 5000), large capitalization stocks are matched exactly in market weights, and then one of the other techniques is applied to find a subset of low capitalization stocks to match the remaining part of the index. For most indexes, because a few stocks (e.g., 20%) make up more than one-half of the market value of the index, this approach has intuitive appeal.

Although an index fund designed to match the S&P index is a popular instrument, managers wanted to make it better almost immediately after it was created. The most obvious way to do this was to find a better proxy for the market portfolio. Index funds exist that track most major indexes such as the New York Stock Exchange (NYSE) index and the Wilshire 5000 stock index. The higher transaction costs of duplicating a broader index mean that one or more of the techniques for matching an index with fewer stocks is almost always used in the construction of a broad-based index fund.

[1]Index funds available to the individual investor, such as the Vanguard funds, maintain cash to accommodate withdrawals. Index funds that invest funds for institutional clients such as pension funds often do not allow withdrawals without substantial notification. In addition, by using dividend reinvestment plans and futures, many are essentially fully invested.

At first thought, one would expect most index funds to underperform the index on average. Index funds have management fees, and transaction costs are incurred in their management. However, two factors help performance. First, S&P occasionally missed small stock dividends in calculating the return on the index, thus understating the actual return. Second, index funds always deliver stock when a firm offers to buy it above market price (in a merger or stock repurchase), but some investors do not. Thus the index fund obtains a higher price for some of its stock than is assumed when the return on the index is calculated. Because of these influences, some low-cost funds have outperformed the index they match over long periods of time.

No index fund has a performance that exactly matches the performance of the index it tracks on a month-by-month or year-by-year basis. Cash inflows from investors, the payment of dividends, and the response to changes in the composition of the index cause the cash position to change and transaction costs to be incurred. Index funds available to individual investors also maintain a cash position to smooth out cash flows. This results in betas slightly below one with respect to the index they track. It also means they generally do slightly better in down markets and slightly worse in up markets. Despite these differences, many index funds earn returns within 0.05% per quarter of the index they track.

The index funds discussed thus far all have a theoretical justification based at least in part on the simple capital asset pricing model (CAPM). We should easily be able to design an index fund based on any of the other equilibrium models described in this book. One type of fund that met with some commercial success was the Wells Fargo yield-tilted index fund. From the posttax equilibrium model developed in Chapter 14, we know that returns should be determined in part by dividends and that the attractiveness of any portfolio to an investor should be a function of the dividend yield on that portfolio and the relative tax rate of the investor.

Thus it makes sense to offer index funds tilted toward (or away from) high-dividend-paying stocks to appeal to investors in different tax brackets. Although these dividend-tilted funds were a success when capital gains and dividends were taxed at different rates, the present tax codes, which tax dividends and realized capital gains at the same rate, have decreased their attractiveness.

The yield-tilted index fund just discussed is one example of a passive portfolio constructed on the basis of a nonstandard CAPM, which is discussed in detail in Chapter 14. It is possible to construct index funds based on any of the other nonstandard CAPM or arbitrage pricing theory (APT) models discussed in Chapters 14 and 16. For example, funds could be constructed with different sensitivities to inflation.

Other types of index funds are based on the kinds of anomalies discussed throughout this book. The investment community soon realized that the Wilshire 5000 tended, over long periods of time, to have a higher return than the S&P 500 (although recently this has not been true). The small stock anomaly was one justification for a more broadly based index. The small stock anomaly (Chapter 17) also has led to the creation of small-stock (low-capitalization) index funds (e.g., funds matching the Wilshire 4500, which excludes the S&P 500, or the Frank Russell 2000, which excludes the top 1,000 stocks). Other new passive management strategies have followed. For example, there are passive portfolios that buy stocks with low price earnings (P/E) ratios. The design of these portfolios depends at least in part on unexplained deviations from theory rather than on theory itself.

Other innovations have also appeared. One popular one uses temporary mispricing across types of markets to increase the return on an index fund. An index fund can hold securities directly or can hold Treasury bills (T-bills) and a future on an index and be in the same risk return position. If futures are underpriced, T-bills plus futures will outperform the index fund that holds stock directly. Some index funds are based on the premise

that futures on average will be underpriced and attempt to outperform holding securities directly by always holding T-bills and futures. Other index funds switch back and forth between the cash and futures markets to take advantage of any temporary mispricing that might arise between the two markets.

Are these passive or active management products? The line of demarcation has blurred. Although the first index funds bought a portfolio of stocks to match the index and did little but reinvest cash, some of the newer funds constantly monitor arbitrage conditions between the cash and futures market and switch between them. There is more of a continuum of products. Where one draws the line between active and passive management is somewhat arbitrary. We have chosen to draw it at the point where forecasts enter the picture. If the manager trades on a mechanical rule by using past data, we call it *passive management*. If the manager forecasts anything and acts on the forecast, we call it *active management*.[2]

ACTIVE MANAGEMENT

Active management involves taking a position different from that which would be held in a passive portfolio, based on a forecast about the future. A decision has to be made about which passive portfolio is best for an investor's goals. For ease of exposition, assume we have made that decision and have decided the appropriate benchmark is the S&P 500 portfolio. The neutral position is to hold each stock in the proportion it represents of the S&P 500. Any difference from these proportions represents a bet based on a forecast. Although there is no universal agreement about classification for active management styles, we find it useful to divide active managers into three groups: market timers, sector selectors, and security selectors. Market timers change the beta on the portfolio according to forecasts of how the market will do. They change the beta on the overall portfolio, either by changing the beta on the equity portfolio (by using options or futures or by swapping securities) or by changing the amount invested in short-term bonds. Although we will have more to say about this in the fourth section of this chapter, market timing in stocks is used far less frequently than market timing decisions in managing bond portfolios.

At the other end of the spectrum from market timing is security selection. The search for undervalued securities and the methods of forming these securities into optimum portfolios have been the subject of much of this book. Investors practicing security selection are betting that the market weights on securities are not the optimum proportion to hold in each security. They increase the weight (make a positive bet) for undervalued securities and decrease it for overvalued securities. Most active stock managers practice security selection.

Another frequently used method of portfolio management is to practice sector or industry selection. This investment style often goes under the name of *sector rotation*. This is

[2]An alternative definition of *passive management* is to only refer to managers who try to replicate an established index as passive managers. This seems to us unduly arbitrary because it would classify differently some managers who do essentially the same thing. As an example, consider the small stock manager. It would include the small stock manager who replicates the Wilshire 4500 stock index. However, it would exclude the small stock manager who buys the lowest decile of stocks in the NYSE. It would also exclude some portfolios such as the Wells Fargo yield-tilted index funds, which clearly were considered passive management by the investment community yet do not replicate an established market index.

Our definition would classify as passive managers those who construct portfolios on the basis of technical analysis. Under our definition, managers who bought on the basis of recent price increases (relative strength) as well as those who bought on low P/E ratios or low capitalization or any other mechanical rule would all be classified as passive managers. We find this more consistent than the convention of the financial community, which considers relative-strength portfolios actively managed and low-capitalization portfolios passively managed. We do not intend by our definition to imply that all styles of passive management are equally good.

like security selection, except that the unit of interest is an industry or a sector. On the basis of analysis, a positive or negative bet will be made on a sector. Although division of the population of stocks by industries is reasonably clear (there is some uncertainty as to how to divide stocks into industries), division by sectors is much more ambiguous.

Firms can divide stocks into sectors in the following ways:

1. broad industrial classification (e.g., industrial, financial, utilities)
2. major product classification (e.g., consumer goods, industrial goods, services)
3. perceived characteristics (e.g., growth, cyclical, stable); other characteristics used to divide stocks into sectors are size, yield, or quality
4. according to sensitivity to basic economic phenomena (e.g., interest-sensitive stocks, stocks sensitive to changes in exchange rates)

The type of analysis under discussion involves selecting one or more of these classifications in more than (less than) market weights according to the anticipated performance. Managers who practice this type of analysis will rotate their portfolios' overweighting (underweighting) sectors or industries over time as they change forecasts of what sector is undervalued or overvalued.

Industry or sector selection should be contrasted to the security selection manager who just picks stocks within one sector. Many managers do not rotate among (select alternative) sectors over time but choose always to select stocks from within one sector or group of industries. These are specialized managers; three examples are growth stock managers, utility stock managers, and technology managers. There are two reasons for specializing. The first is the belief that the sector is permanently undervalued. The second is the belief that one's staff is better able to select undervalued stocks in that sector or industry than in any other. Although the first is hard to justify, the second can be justified in a world of increasing complexity and specialization. The client can invest in sectors that are not covered by a particular manager by using either other active managers or a passive portfolio that covers these sectors (often called a *completion fund*).

PASSIVE VERSUS ACTIVE

The case for passive versus active management certainly will not be settled during the life of the present edition of the book, if ever. Active management has some costs to overcome if it is to be effective. The predictive content of forecasts used in active management must be sufficiently large to overcome the following costs:

1. The cost of paying the forecasters either in the form of salaries or in the higher management fees charged by active managers relative to passive managers.[3]
2. The cost of diversifiable risk. Active portfolios, by their nature, have more diversifiable risk than an index fund (which has close to zero). The investor must be compensated for taking this risk.[4]

[3]In 2010, Vanguard's index fund had an expense ratio of 0.12%, whereas the active average fund had an expense ratio of 125%, and Vanguard's index fund had an 8% turnover of assets compared with the turnover for a typical stock mutual fund of 90%.

[4]As an approximation to this, consider active managed stock mutual fund data. First mutual funds have coefficient of determination (R^2), with the market of between 0.90 and 0.95. To have the same total risk, an S&P index fund would need a beta of 1.025–1.050. If the risk premium in the market is 6%, this implies that active funds would need to earn an added return of between 0.15% and 0.30% to compensate for their added risk.

3. The cost of higher transaction cost. Active decisions require turnover as opposed to the very low turnover of the buy and hold strategies of an index fund.
4. For the taxable investor, an early incidence of capital gains tax. Under current tax laws capital gains or losses are realized for tax purposes either because the fund sells stocks or the investor sells all or part of his share in the fund. An index fund has a very low level of turnover, so the taxable investor pays minimal capital gains taxes until he sells off part of the fund. An actively managed portfolio usually has a much higher turnover, and so capital gains taxes can be incurred by the investors even when the investors wish to leave their money fully invested.

Although index funds have outperformed most active managers, most investors who hire active managers believe they can spot the manager who will outperform the index. This belief persists despite the fact that there is very little evidence that superior performance is predictable. We are reminded of a recent survey of the entering class of one of the country's top-rated colleges. When students were asked if they expected to finish in the top 10% of their class, 87.5% responded that they did.

INTERNATIONAL DIVERSIFICATION

At several points in this book we have discussed the potential of international portfolio diversification. Although European portfolio managers have regularly diversified internationally, implementation of the concept is a much newer concept to American portfolio managers. Once again, the portfolio manager has a choice involving passive and active management. At the most aggregate level, the manager must decide whether to hold a passive portfolio of individual countries, weighting perhaps by the aggregate market value of each portfolio, or to try to select undervalued countries. Second, the manager must decide whether to bear the potential benefits and risk of currency movements or to hedge away changes in the relative value of currencies.

Once this decision is made, the manager must decide whether to actively or passively manage the portfolio of stocks within each country. Any of the models discussed to this point can be applied within each country.

BOND MANAGEMENT

Many of the portfolio management strategies that are utilized for common stocks are also utilized for bonds. Although we will briefly review the justification for these strategies when applied to bond portfolio management, we will concentrate on those strategies that are unique to the bond area. Once again it is convenient to divide the strategies into passive and active strategies. In this section we will discuss strategies where the manager is concerned only with the return characteristics of the bond portfolio. In the next section we will discuss bond strategies that consider the existence of liabilities, such as immunization strategies.

Passive Strategies

As in common stock portfolio management, one passive strategy for bonds is to match an index. An institutional investor estimated that $67 billion of institutional assets was invested in domestic bond index funds and $1 billion in international index funds as of 1990. Index matching for common stocks is justified primarily by the equilibrium arguments of Chapters 13 and 14. Although some equilibrium models for bonds have been

developed, empirical testing of the existing theories is virtually nonexistent. The justification for bond index funds rests primarily on the performance of active versus passive portfolios rather than on tests of a theory. Performance statistics reported in the industry press generally find that most active bond managers underperform the generally utilized bond indexes, such as the Shearson–Lehman index and the Salomon Brothers index.[5] If this past performance continues in the future, an investor could expect above average performance by investing in an index fund.

Some features of bond index funds differ from stock index funds and make bond index funds more difficult to manage. The first factor is the changing nature of the index. Stock indexes, such as the S&P 500, change occasionally as the S&P decides that different firms are more appropriate or as firms in the index merge. The composition of all widely used bond indexes changes much more frequently as bonds mature and new bonds are issued. A second difference from stock indexes is that many bond indexes contain bonds that are illiquid and, in fact, might not be available to an investor. This means that the manager of a bond index fund will never attempt to duplicate an index exactly but will employ one of the other techniques described in the section on stock index funds to match the index. Both of these influences mean that the passive manager will have to trade more frequently in running a bond index fund than she does in running a stock index fund.

Another type of passive product is unique to the bond area—the bond unit trust usually composed of municipal bonds. A unit trust buys a portfolio of bonds and does not buy or sell bonds over the life of the trust. A trust unwinds gradually over time and eventually dissolves as the bonds it initially buys are called or mature. An investor purchases a share of this fixed portfolio. Ignoring default and calls for the moment, the investor knows the cash flows associated with the portfolio and the bonds that comprise the portfolio. The important question is what benefits the unit trust provides that are not provided by other opportunities.

Investors could simply duplicate the holdings of the unit trust on their own and, by doing so, avoid management and sales fees. Investors of modest means would be faced with the problem that bonds can be bought in only large denominations (are not divisible) and that transaction costs of buying small amounts are high. Thus duplication is not feasible for most investors.

Two other opportunities suggest themselves—holding other bonds or index funds. A unit trust has the advantage of diversifying risk rather than holding a small number of bonds directly. In addition, since for most bonds the largest payment occurs when the principal is repaid, a portfolio of bonds with varying maturities can offer a more uniform cash flow. For retired people who want to consume capital as well as interest over time, the more uniform cash flow offered by a unit trust is often viewed as desirable.

The other comparison is with an index fund. A unit trust offers many of the benefits of diversifying default risks and, to a limited extent, call risks that an index fund offers.[6] However, it offers more predictable cash flows and larger cash flows in the early years, because a unit trust pays out principal as bonds mature or are called, whereas an index fund

[5]Standard indexes are market weighted and are therefore dominated by short-term bonds. The indexes have an average duration of three to five years. Bond managers generally hold portfolios of longer duration; thus performance of funds relative to the index is strongly affected by how interest rates changed during the period in which funds are being examined.

[6]The principal determinant of calls is the level of interest rates. The effect of the level of interest rates on calls is common to all bonds and is not diversifiable. However, bonds are called for other reasons such as firm restructuring. This part of the call risk can be diversified across bonds.

reinvests principal. Of course, this faster payment of principal can be an advantage or a disadvantage according to whether the investor wants money back earlier.[7] Unit trusts are the major passive strategy unique to bonds. Let us now examine active strategies.

Active Strategies

Active bond strategies are very similar to active stock strategies, although the popularity of each of the strategies differs. The most commonly employed active bond strategy is market timing. An estimate is made of what will happen to interest rates. If interest rates are expected to rise, prices are expected to fall, and capital losses will be incurred on bonds. Thus an investor expecting a rise in interest rates will shorten the duration of the portfolio, while an investor expecting a drop in interest rates will lengthen the duration of the portfolio. Although this is the standard argument underlying market timing, the argument is incomplete, and additional comments are in order.

If a manager believes that interest rates will rise, and other investors share this belief, then market prices will reflect this expectation. In this situation, market timing will be ineffective, even with correct forecasting. Thus successful market timing requires both accuracy in forecasting and beliefs different from those already reflected in market prices.

A second active bond strategy is to pursue the risk premium associated with lower-rated bonds. (This is one form of sector selection discussed in the section on common stocks.) Evidence discussed in Chapter 21 suggests that the extra promised return on riskier corporate bonds has historically more than compensated for the loss due to default. For example, AAA corporates have minimal default risk and yet over the last several decades have offered a promised return well above similar government bonds. The same evidence indicates that historically, the extra promised return of low-rated corporate debt relative to high-rated corporate debt more than compensated for the default losses.

A strategy followed by some portfolio managers is to try to earn an extra return by bearing credit risk. By holding a sufficiently large portfolio, the probability of a large portion of the portfolio defaulting is small, and with a positive risk premium, an extra return is earned. The manager of a high-risk portfolio will usually try to improve performance by using fundamental analysis to screen out the bonds that are most likely to default and to identify those most likely to show improvement in creditworthiness.

The risk of this strategy is twofold. First, the manager may have higher default experience than is expected. As shown in Chapter 21, default experience varies a great deal on a year-to-year basis and is related to current economic conditions. A manager using long-term experience as a guide to estimating the extra return from bearing credit risk might experience defaults in excess of historical experience and may have realized returns below those on higher-rated debt. The second risk involved in investing in lower-rated debt is a change in the premium. For example, the difference between AAA corporate debt and government debt with similar characteristics might be 0.3%. Historical experience would suggest that the default possibilities for AAA corporate debt are exceedingly low. Thus holding AAA corporate debt rather than government debt would seem to almost guarantee an extra return. There have been a number of periods, however, when the realized return of government debt exceeded that of AAA corporate debt. How can this occur? If the yield spread between AAA corporates and governments widens from, for example, 0.3% to 0.5%, then

[7]The investor holding an index fund can, of course, obtain larger cash flows in early years by selling off part of the fund periodically. However, the size of cash flows is less predictable because of changes in the market value of the fund's portfolio.

AAA corporates will experience a capital loss relative to governments. This is the second risk in pursuing a lower-rated debt strategy.

The previous discussion suggests a third active strategy analogous to the strategy of sector rotation discussed for stocks. If a manager anticipates that the spread between AAA corporates and governments would widen significantly, then a switch from AAA corporates to governments should result in a better performing portfolio. Similarly, if the spread is expected to narrow or to remain unchanged, AAA corporates should have the superior returns. This strategy can be classified as sector selection. A category such as AAA corporates is selected based on the belief that this sector will have superior performance.

Although selecting a sector by comparing relative spreads between bonds of different rating category is the most obvious example of sector selection, the principle is perfectly general. Callable debt has a higher yield than noncallable debt because of the risk that the issuer will call debt at a disadvantageous time. Callable debt can be viewed by an investor as purchasing noncallable debt and issuing a call option (which reduces the value of the bond).

If the investor believes the market overprices the option (overestimates the probability of a call), then purchasing the callable debt should lead to superior returns. Both rating category and call features are examples of potential sector selection.

Security selection as a strategy in the bond area is the same as in the stock area; however, there is much less chance for excess returns. Security selection of bonds generally involves one of two approaches. One approach is to search for securities whose default risk is misestimated. For example, are there firms with A-rated bonds that have substantially less default risk than other A bonds (have been misclassified)? This strategy involves credit analysis. A second approach is to try to find bonds that are mispriced given their characteristics. A number of commercial services estimate the "fair" yield to maturity on a bond given its characteristics (bond rating, maturity, callability, etc.); Barra and Gifford Fong are two examples of such services. Buying bonds whose actual yield to maturity differs from model yield is a security selection strategy. The difference in model and actual yield is generally quite small. Thus the extra return is also quite small, even if actual yield moves to model yield. In the common equity area, finding a stock of a high-growth company before the market recognizes it can lead to spectacular returns. Thus security selection strategies in the common equity area have the potential for much higher returns.

We have discussed the major strategies in bond management when the investment manager is concerned only with returns. When the investor is also concerned with a liability stream, the strategies change considerably.

BOND AND STOCK INVESTMENT WITH A LIABILITY STREAM

Many portfolio managers are in charge of investing funds that are provided to meet future obligations. Managers of pension funds are the most obvious example; managers of insurance companies are another. There has been a greater awareness in recent years that the portfolio manager needs to consider the liability stream in making investment decisions. There are several reasons for this increased awareness. First, the accounting treatment of the return on pension assets has changed. This change means that changes in asset values relative to liabilities affect the earnings a company reports to its shareholders and affect the asset and liability values shown on the balance sheet. Second, regulatory bodies concerned with financial intermediaries such as insurance companies have forced the intermediaries to value more of their assets and liabilities at the price they would get if sold rather than at original cost. Third, in the 1980s, many companies found that certain investment strategies, such as cash flow matching or dedication, resulted in surplus pension assets, and thus

pension assets became a source of money.[8] All of these factors led to a greater awareness of the need to consider liabilities in selecting investment strategies.

When an investment manager considers the cash flow characteristics of liabilities as well as those of assets, investment strategies change. There are two ways to model the liability stream that are useful for formulating investment strategies. One is to assume that the liability stream is known and fixed; the other is to assume that the liability stream is a function of one or more exogenous influences. Each of these is now discussed in some detail.

Fixed Liability Stream

A manager is often called on to manage a portfolio of bonds so as to meet a fixed set of liability payments over time. Although liabilities often are not truly fixed, there are circumstances in which acting as if they were fixed is a close approximation to reality. Probably the clearest case was the sale of guaranteed insurance contracts by insurance companies, which were very popular in the early 1980s and still are. These contracts required that the insurance company pay a fixed sum at specified intervals to the purchaser. The contracted payments were fixed commitments (liabilities) of the insurance company, which had to put together portfolios of bonds so that it could meet these payments.

Another example is pension payments for retired employees. If the pension payments are fixed at the time of retirement, the amount per year that must be paid to any employee is fixed. Of course, the aggregate amount paid to all employees is not known because the mortality experience of the employees is not known. However, with a large number of retired employees the mortality experience can be predicted quite accurately. Hence pension funds frequently require an investment manager to protect against a set of fixed forecasted pension liabilities. Low-risk strategies for managing bond portfolios to meet a fixed liability stream were discussed in Chapter 22. These include cash flow matching (often called dedication) and immunization.

An exact cash flow–matched portfolio employing only noncallable default-free (government) debt would have zero risk. Managers of cash flow–matched or dedicated portfolios are often selected on the basis of the initial value of the assets they require to meet the pension liabilities of a client. Managers in bidding on business wish to be competitive, and they can increase return (decrease initial assets) in two ways. The first, already discussed in Chapter 22, is to allow cash to be transferred between periods at a set rate. This is an assumed rate, and to the extent that it does not materialize, liabilities will not be matched.

The second is to introduce higher expected return debt into the portfolio. Managers frequently use corporate debt or callable debt to raise the expected return (lower the initial cash) necessary to match a set of liabilities. Although this might allow the manager to gain a customer, it increases the probability the liabilities will not be met. A corporate bankruptcy or an early call after a drop in interest rates will result in a shortfall in the return produced by assets.

The next most risky bond strategy is immunization. Because the liability stream is fixed in amount and timing, the only uncertainty involved in determining its value is the appropriate discount rates for valuing it. As the yield structure changes, the present value of the

[8]This was a major source of funds used in many leveraged buyouts. Surplus pension assets were recognized because actuaries were willing to value liabilities differently, depending on the investment strategy. If the firm's cash flow matched pension assets, the actuary used the return on the investment portfolio as the discount rate in valuing liabilities. This was a much higher rate than they normally used and resulted in a lower value of liability and excess pension assets.

liabilities also changes. If the investment manager is to have assets of at least equal value to the liabilities at all points in time, the manager will need to have the assets change in value in the same manner as the liabilities.

Because the value of the liabilities is dependent only on the term structure, the element in the investment policy that will affect risk is the sensitivity of assets to changes in interest rates. A policy of having the assets have the same sensitivity to interest rates as the liability stream (called an immunized policy) is a low-risk strategy. An immunized strategy is higher risk than cash flow matching because it depends on the accuracy of the measurement of the sensitivity of bonds to a change in interest rates.[9]

An immunized strategy can have a large component of active management. All of the active bond portfolio management techniques discussed in the previous section can be used in conjunction with immunization; these include sector selection and security selection. The manager maximizes the extra return subject to the constraint that the portfolio is immunized.[10]

Some managers attempt to do a modest amount of market timing while maintaining an immunized portfolio on average. A manager who wishes to market time would deviate from an immunized policy in some periods. Assume for a moment that duration is a reasonable measure of sensitivity to interest rate changes. If the manager believed that interest rates will rise, then a negative net duration would be set on the portfolio (duration on assets less than duration on liabilities). Similarly, if interest rates are expected to fall, a positive net duration would be set.

As a manager introduces more elements of active management into immunization strategies, she is attempting to increase expected returns but, in doing so, is increasing the probability that liability payments will not be met. In short, we are back to a risk–return choice, although a very restricted one. For example, because of their low correlation with interest rate movements, common stocks are not likely to be effective in immunization. Thus a manager investing in common stocks and concerned with a fixed liability stream has engaged in high-risk strategy in an attempt to earn a high return.

The choice of how immunized a manager should be depends in part on the ratio of assets to liabilities and in part on how the assets are funded. For example, a typical type of manager who practices immunization is the manager of a large financial intermediary, such as an insurance company. Many of these institutions have assets that are only slightly larger than their liabilities. An insurance company might have $20 billion in assets, $19.5 billion in liabilities, and $0.5 billion in net worth. For such an institution, a policy that results in a small fluctuation in asset value without a corresponding change in liabilities can be disastrous. In the example, more than a 5% decline in asset value without a decrease in the value of liabilities results in negative net worth. Thus, for these types of highly levered institutions, a cash flow–matched or immunized bond investment policy is the only reasonable policy. In other circumstances a higher return investment strategy is more reasonable. For example, a company financing a pension plan for retired employees may be willing to risk a decline in assets to below liabilities because of the chance of a substantially higher return. Another possibility is an investment strategy called *contingent immunization*. The manager is active until the value of the assets just equals the value of the liabilities (or) the value of

[9]A cash flow–match strategy is, of course, also immunized. However, it is useful to distinguish those portfolios that are immunized, because they are cash flow matched, and those that are immunized by matching sensitivities to interest rates.

[10]There may be a cost to requiring the portfolio to be immunized when the manager believes that some bonds are mispriced. In this case there is a risk–return trade-off. The manager may choose to incur the risk of a nonimmunized portfolio to make a heavier investment in bonds that are believed to yield an excess return.

the liabilities plus a fixed amount. At this point the active portfolio is liquidated and the portfolio is immunized. This strategy allows active management but has the guarantee of immunization.

When the liability stream is stochastic rather than fixed, management techniques become more complex. This is the situation that we now examine.

Stochastic Liability Stream

Stochastic liability streams arise in a number of realistic investment situations. Retired employees often have a cost-of-living adjustment (COLA) in their retirement plans. Their base pension is fixed but is adjusted upward because of changes in some index such as the consumer price index. Casualty insurance (such as automobile insurance) companies are another example. A casualty insurance company receives premium income. The size of the liabilities (e.g., auto accident claims) takes a number of years to determine because of the time needed to litigate claims. Although the number of claims can be estimated fairly accurately, the size of each claim is uncertain. The ultimate settlement is generally assumed to be a function of medical costs and general living costs. The assumption of a stochastic liability stream with a fixed component and a variable component related to medical costs and inflation is reasonable for many casualty insurance companies.

To provide a concrete basis for subsequent discussion, let us consider the investment problem of a manager who is concerned with providing pension funds for retired employees whose benefits include a COLA. Without the COLA the only uncertainty in determining the present value of the liabilities is the appropriate discount rate to use. The rate of discount depends on the term structure of interest rates. Thus the rate of change in the value of liabilities is affected by shifts in the term structure. With a COLA the rate of change in liabilities is affected by two factors: changes in the term structure and the rate of inflation. We can then think of rates of change in the liabilities as being determined by a two-factor model.

If a manager is concerned only with variability of return on assets, then the sensitivity of the assets to a factor determines risk. When a manager is investing subject to a liability stream, the risk of movements in a factor is eliminated by making the net exposure (asset exposure less liability exposure) zero.

Previously, we discussed immunization as a strategy. This involves eliminating exposure to shifts in the term structure by setting the net exposure to shifts in the term structure to zero. With the two-factor model, the manager can choose to eliminate the exposure to any factor by making the net exposure to that factor zero.

Should a manager eliminate factor risk? To answer this question, consider two different scenarios. First, consider that the factor is unpriced so that exposure does not affect expected return in equilibrium. Second, consider that the factors are priced in the sense that exposure increases expected return.

If the factor is unpriced, then exposure to the factor in equilibrium produces extra risk without any additional expected return. If the manager does not have a special ability to forecast period-by-period values of the factor, zero net exposure (immunization) is the preferred strategy. A belief by the manager in an ability to forecast the period-by-period value of the factor can mean a nonzero and changing exposure through time.

If the factor is priced, then the manager might choose to have exposure to the factor. Exposure to a factor increases the risk but also potentially increases the return. Thus, for priced factors, exposure to the factor involves a risk–return trade-off.

Nothing in our prior discussion requires that the assets be only bonds. The inflation factor for assets (and liabilities) is the effect of inflation holding changes in interest rates

constant. For many bonds most of the effect of inflation will be impounded in a change in interest rate. Thus investment managers concerned with a liability stream whose value is affected by inflation as well as interest rate changes might find both common equities and bonds to be useful hedging tools.

What is true of inflation is also true of other factors affecting the liability stream. Noninterest factors are generally better immunized by using common stocks. Return-generating models for common stocks tend to have many more components than bond return-generating models. Furthermore, many liability streams are likely to be sensitive to many of the additional influences that drive stock returns. Thus the incorporation of stocks in the asset portfolio generally allows better hedging of noninterest rate factors in the liability stream.[11]

The concept of using a multi-index return-generating process to immunize a set of liabilities is new, so very little theoretical or empirical research exists in this area.[12] We expect it to be a fruitful area for future research.

Bond–Stock Mix

Management style with respect to bond–stock mix can be divided into two broad categories: managers who use fairly stable proportions through time and managers who actively vary their proportions over time.

The managers who use fixed proportions generally make an assumption that the characteristics of various asset classes (e.g., mean return, variance) are fairly constant over time, or at least that changes in these parameters are not predictable. They generally examine the distribution of various combinations of asset classes and decide which combination is the most attractive. Often managers are concerned with the return characteristics over several multiperiod time horizons. In this case, simulation is frequently used to examine the characteristics of the return distributions for different mixes over several different time spans. Frequently, assumptions concerning asset characteristics are then varied to determine whether the chosen mix is reasonable with small changes in assumed values of the return distributions.[13]

Two management styles lead to varying mixes over time. The first is the market timer. Some managers believe that they can forecast the relative performance of the stock and bond market. The current term for this is *tactical asset allocation*. These managers have a variety of techniques for forecasting relative performance, which vary from a mechanical rule for utilizing past data to reliance upon forecasted changes in risk–return relationships.

One justification for this behavior is the empirical literature supporting changing risk premiums. In a series of papers, Fama and French (1987) have shown that for holding periods beyond a year, expected returns on stocks and long-term bonds and, hence, relative returns are weakly predictable. They present results that variables such as dividend price ratios, default premiums, and term premiums can explain more than 30% of the variation

[11]An example of a more complex liability stream is the forecasted liability stream associated with a pension plan for active employees. The value of this stream is likely to depend on interest rates, cost-of-living changes, changes in the risk of the company and economy, changes in profitability, and so forth. Some of these influences affect the return on bonds, but more affect the return on stocks.

[12]See Elton and Gruber (1990, 1991, 1992).

[13]Another condition necessary for fixed proportions is that the investor's utility function is not one that results in optimal asset proportions being a function of the current value of the portfolio (wealth). Managers who utilize fixed proportions either do not think in terms of utility functions or are explicitly making the assumption just discussed.

in returns. Among those who provide similar evidence are Keim and Stambaugh (1986), Poteba and Summers (1988), and Campbell and Shiller (1987).

A theoretical justification for changing asset proportions is that the behavior of investors can be characterized by a utility function that implies optimal asset proportions are a function of the value of the portfolio. If the utility function depends on the value of the portfolio, the optimal asset mix at any point in time depends on the returns in all previous periods. Several firms use an explicit utility function to select the optimal bond–stock mix over time.

The second management style leading to changing asset proportions over time is one that uses changes in the asset mix to change the shape of the return distribution. As discussed in Chapter 23, a changing mix of stocks and T-bills can replicate the pattern of holding the stock portfolio along with a put on the stock portfolio. This trading strategy often goes under the name of *portfolio insurance* or *dynamic asset allocation*. Leland, O'Brian, and Rubinstein were leading proponents of this type of product.

Any manager who chooses to change the bond–stock mix over time can accomplish this goal through transactions in the futures market as opposed to the cash market. To increase the exposure to stocks relative to bonds, the manager has to buy stock futures and sell bond futures. The use of futures has great appeal because of the low transaction costs and high liquidity in the futures market as compared with the cash market. Futures also allow the bond–stock mix to be easily modified without changing the exposure to specific issues that the manager may wish to maintain. This characteristic is especially useful for pension managers who employ multiple managers. Implementing the bond–stock choice by utilizing futures allows the bond–stock decision to be controlled at the level of the aggregate portfolio. This means that each of the multiple managers need not be concerned with the bond–stock mix and can manage assets in a segment or segments of the financial markets that she believes is appropriate. Finally, if futures are used to implement timing decisions, it becomes easy to separate the return from timing decisions from the return due to selection ability.

BIBLIOGRAPHY

1. Admati, Anat R. "Does It All Add Up? Benchmarks and the Compensation of Active Portfolio Managers," *Journal of Business*, **70,** No. 3 (July 1997), pp. 323–350.
2. Campbell, John Y., and Shiller, Robert. "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," unpublished manuscript, Princeton University (1987).
3. Chevalier, Judith. "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, **105,** No. 6 (Dec. 1997), pp. 1167–1200.
4. ——. "Career Concerns of Mutual Fund Managers," *Quarterly Journal of Economics*, **114,** No. 2 (May 1999), pp. 389–443.
5. Cumby, Robert E., and Glen, Jack D. "Evaluating the Performance of International Mutual Funds," *Journal of Finance*, **45,** No. 2 (June 1990), pp. 497–521.
6. Elton, Edwin J., and Gruber, Martin J. "A Multi-index Risk Model of the Japanese Stock Market," *Japan and the World Economy*, **1,** No. 1 (1989), pp. 21–44.
7. ——. "Expectational Data and Japanese Stock Prices," *Japan and the World Economy*, **1,** No. 4 (1990), pp. 391–401.
8. ——. "Optimal Investment Strategies with Investor Liabilities," *Journal of Banking and Finance*, **10,** No. 2 (March 1991), pp. 210–230.
9. ——. "Portfolio Analysis with a Non-normal Multi-index Return Generating Process," *Review of Quantitative Finance and Accounting*, **2,** No. 1 (March 1992), pp. 100–120.
10. Fama, Eugene F., and French, Kenneth R. "Dividend Yields and Expected Stock Returns," unpublished manuscript, University of Chicago (1987).

11. French, Kenneth R., Schwert, G. William, and Stambaugh, Robert. "Expected Stock Returns and Volatility," *Journal of Financial Economics*, **19** (1986), pp. 3–30.
12. Grinblatt, Mark, Titman, Sheridan, and Wermers, Russ. "Momentum Investment Strategies, Portfolio Performance, and Herding: A Study of Mutual Fund Behavior," *American Economic Review*, **85,** No. 5 (Dec. 1995), pp. 1088–1105.
13. Gruber, Martin J. "Another Puzzle: The Growth in Actively Managed Mutual Funds," *Journal of Finance*, **51,** No. 3 (July 1996), pp. 783–810.
14. Hendricks, Darryll, Patel, Jayendu, and Zeckhauser, Richard. "Hot Hands in Mutual Funds: Short-Run Persistence of Relative Performance, 1974–1988," *Journal of Finance*, **48,** No. 1 (March 1993), pp. 93–130.
15. Ippolito, Richard A. "Efficiency with Costly Information: A Study of Mutual Funds," *Quarterly Journal of Economics*, **104,** No. 1 (Feb. 1989), pp. 1–21.
16. Keim, Donald B., and Stambaugh, Robert F. "Predicting Returns in the Stock and Bond Markets," *Journal of Financial Economics*, **17** (Dec. 1986), pp. 357–390.
17. Khorana, Ajay. "Top Management Turnover: An Empirical Investigation of Mutual Fund Managers," *Journal of Financial Economics*, **40,** No. 3 (March 1996), pp. 403–427.
18. Narayanan, M. P. "Form of Compensation and Managerial Decision Horizon," *Journal of Financial and Quantitative Analysis*, **31,** No. 4 (Dec. 1996), pp. 467–491.
19. Poteba, James M., and Summers, Lawrence H. "Mean Reversion in Stock Prices," *Journal of Financial Economics*, **21** (Oct. 1988), pp. 27–59.
20. Wahal, Sunil. "Pension Fund Activism and Firm Performance," *Journal of Financial and Quantitative Analysis*, **31,** No. 1 (March 1996), pp. 1–23.

Index