Regression analysis

The term regression analysis refers to the methods used to estimate the values of a variable from a knowledge of the values of another variable. Regression analysis is a statistical technique for investigating and modeling the relationship between variables. The reason why it is so widely applied is because it provides the answer to an everyday question, namely **how a response variable of special interest depends on several other, explanatory variables.**

In various quantitative settings, the regression techniques models the relationship between the response variable of special interests (Y) and a (set) x1, . . . xk of explanatory or predictor variables.

Linking the response variable to the predictors:

$$Y = \underbrace{f(x_1, \ldots, x_k)}_{\text{deterministic}} + \underbrace{\varepsilon}_{\text{random}},$$

ε is the error term which is can neither be controlled or predicted. The goal is to learn about the function f (·). In full generality, finding f (·) without any conditions is very difficult: function space is infinite-dimensional.

Solution: restrict the form of f (·).

Linear modeling: Y = β0 + β1x1 + · · · + βk xk + ε.

Finding f boils down to determining β0, β1, . . . , βk from the data.

Types of regression analysis

It is classified into many types according to the type of relationship between the two variables such as

- Linear
- Exponential
- Logarithmic
- Power regression analysis

Type of Regression	Conditions
Univariate	Only one quantitative response variable
Multivariate	Two or more quantitative response variables
Simple	Only one explanatory variable
Multiple	Two or more explanatory variables
Linear	All parameters enter the equation linearly, possibly after transformation of the data
Nonlinear	The relationship between the response and some of the explanatory variables is nonlinear or some of the parameters appear nonlinearly, but no transformation is possible to make the parameters appear linearly
Analysis of variance	All explanatory variables are qualitative variables
Analysis of Covariance	Some explanatory variables are quantitative variables and others are qualitative variables
Logistic	The response variable is qualitative

Simple linear regression

Simple linear regression is used to estimate the relationship between **two quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g., the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g., the amount of soil erosion at a certain level of rainfall).

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Simple linear regression example: You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from 15k to 75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

If you have more than one independent variable, use multiple linear regression instead.

Assumptions of simple linear regression

Simple linear regression is a **parametric test**, meaning that it makes certain assumptions about the data. These assumptions are:

1. **Homogeneity of variance (homoscedasticity)**: the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. **Independence of observations**: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. **Normality**: The data follows a normal distribution.

Linear regression makes one additional assumption:

4. The relationship between the independent and dependent variable is **linear**: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

If your data do not meet the assumptions of homoscedasticity or normality, you may be able to use a nonparametric test instead, such as the Spearman rank test.

Example: Data that doesn't meet the assumptions. You think there is a linear relationship between cured meat consumption and the incidence of colorectal cancer in the U.S. However, you find that much more data has been collected at high rates of meat consumption than at low rates of meat consumption, with the result that there is much more variation in the estimate of cancer rates at the low range than at the high range. Because the data violate the assumption of homoscedasticity, it doesn't work for regression, but you perform a Spearman rank test instead.

If your data violate the assumption of independence of observations (e.g., if observations are repeated over time), you may be able to perform a linear mixed-effects model that accounts for the additional structure in the data.

How to perform a simple linear regression

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- $\mathbf{B_0}$ is the **intercept**, the predicted value of **y** when the **x** is 0.
- $\mathbf{B_1}$ is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B_1) that minimizes the total error (e) of the model.

While you can perform a linear regression by hand, this is a tedious process, so most people use statistical programs to help them quickly analyze the data.

Multiple linear regression

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Multiple linear regression is used to estimate the relationship between **two or more independent variables** and **one dependent variable**. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Multiple linear regression example: You are a public health researcher interested in social factors that influence heart disease. You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship between them.

Assumptions of multiple linear regression

Multiple linear regression makes all of the same assumptions as simple linear regression:

Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.

Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among variables.

In multiple linear regression, it is possible that some of the independent variables are actually correlated with one another, so it is important to check these before developing the regression model. If two independent variables are too highly correlated ($r2 > \sim 0.6$), then only one of them should be used in the regression model.

Normality: The data follows a normal distribution.

Linearity: the line of best fit through the data points is a straight line, rather than a curve or some sort of grouping factor.

How to perform a multiple linear regression

Multiple linear regression formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The t statistic of the overall model.
- The associated p value (how likely it is that the t statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the t statistic and p value for each regression coefficient in the model.

Between Correlation and Regression

Correlation	**Regression**
'Correlation', as the name says, it determines the interconnection or a co-relationship between the variables.	'Regression' explains how an independent variable is numerically associated with the dependent variable.
In Correlation, both the independent and dependent values have no difference.	However, in Regression, both the dependent and independent variables are different.
The primary objective of Correlation is to find out a quantitative/numerical value expressing the association between the values.	Regression's main purpose is to calculate the values of a random variable based on the values of a fixed variable.

Correlation stipulates the degree to which both variables can move together.	However, regression specifies the effect of the change in the unit in the known variable(p) on the evaluated variable (q).
Correlation helps to constitute the connection between the two variables.	Regression helps in estimating a variable's value based on another given value.