Regression Analysis

Munmun Akter

Lecturer

NFE, DIU

What is Regression analysis?

- A statistical technique used to **examine the relationship** between one or more **independent variables (predictors)** and a **dependent variable (outcome).**
- It is commonly used for various purposes, including:

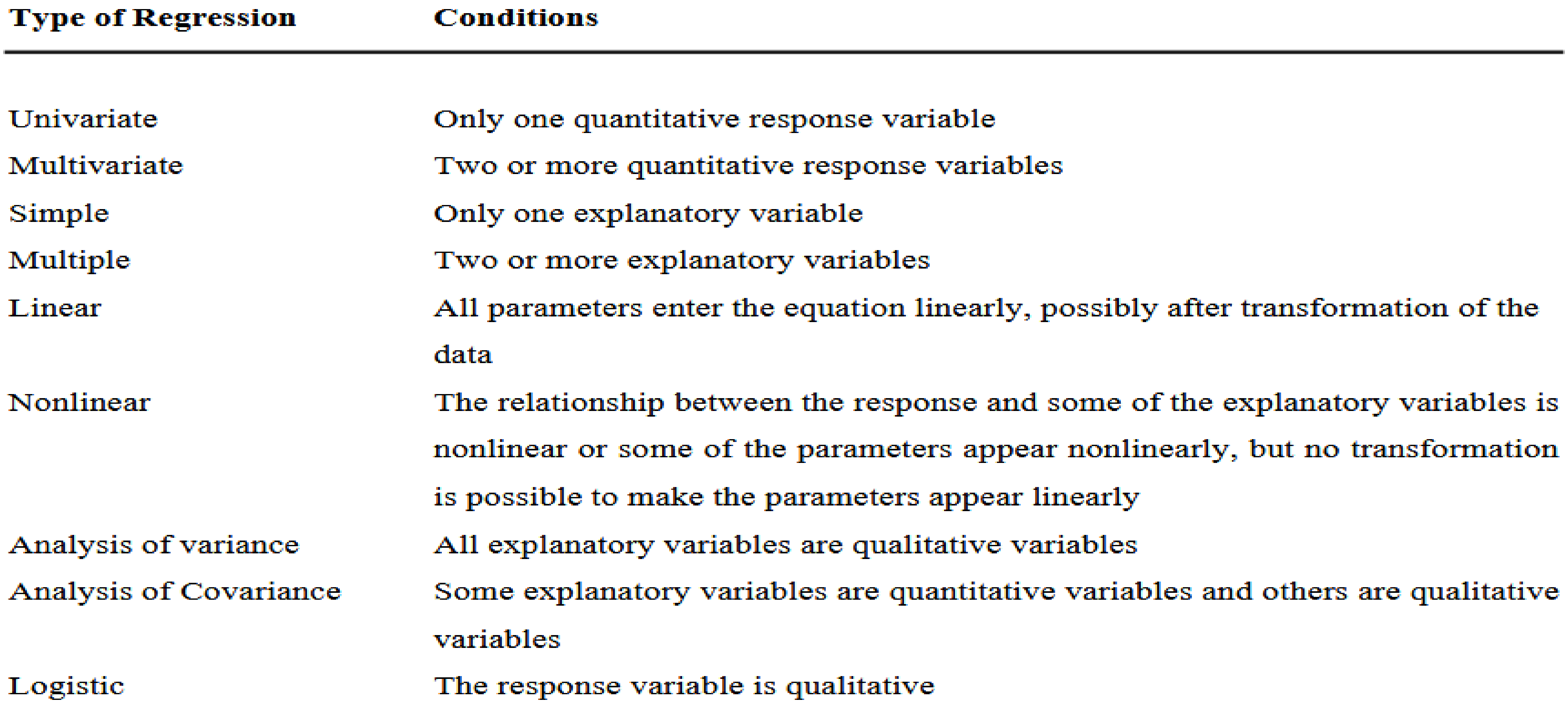
1. **Prediction**: For example, you might want to predict a person's salary based on their education, years of experience, and other factors.
2. **Causation**: Doesn't prove causation, can provide the **strength and direction** of relationships between variables.
3. **Modeling**: Used to create mathematical models for **simulations and scenario analysis.**

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

Types of regression analysis

- Based on different methodologies:



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 Liner Regression

used to model the **relationship between two variables**: a **dependent** variable (often denoted as "Y") and an **independent** variable (often denoted as "X"). It is called "**simple**" because it deals with the relationship between **just two variables**, as opposed to **multiple variables** in multiple linear regression.

The linear equation

$$\mathbf{Y = a + bX}$$

Y represents the **dependent variable** (the one you want to predict or explain).

X represents the **independent variable** (the one you **use to make predictions**).

a is the **intercept**

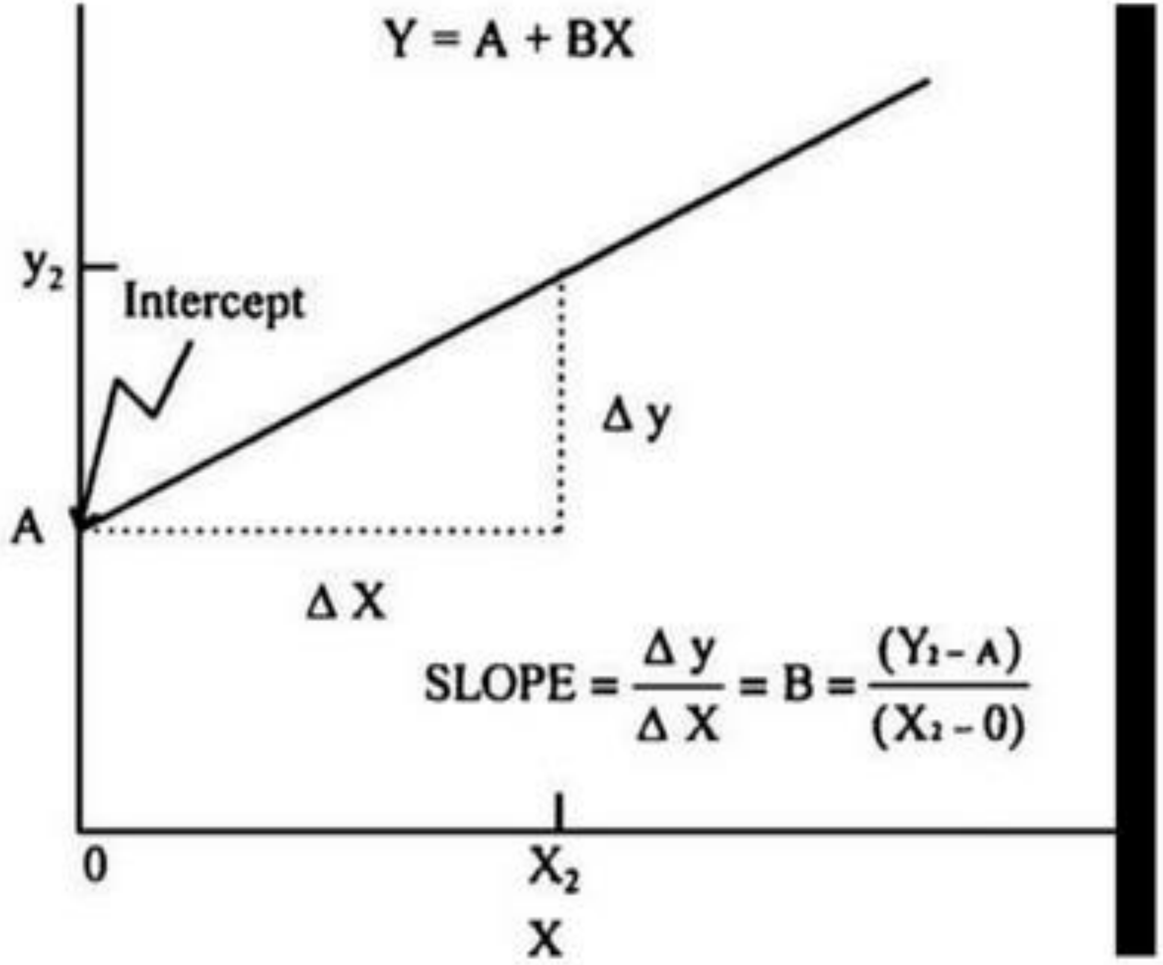
b is the **slope**, which represents the change in Y for a one-unit change in X.

- The **primary objective** of simple linear regression is **to estimate the values of a and b** that **best fit the observed data points**.
- The values of **a and b** are usually estimated using techniques like the **Method of least squares.**
- **Slope = b = $(X-X)(y-y)/ \Sigma (X-X)^2$**

  Or, b=$(N\ \Sigma Xy)-(\Sigma X)(\Sigma y)/(N\ \Sigma X^2-(\Sigma X)^2)$
- **Intercept = a = mean y-b* mean X**

Or, a= $\Sigma y/N$

Straight-line plot.


Y = A + BX
y_2
Intercept
Δy
A
ΔX
SLOPE = $\frac{\Delta y}{\Delta X}$ = B = $\frac{(Y_2 - A)}{(X_2 - 0)}$
0
X_2
X

Calculation of Simple Linear Regression

We'll use the following dataset:

X (Hours Studied)	Y (Test Scores)
2	65
3	75
4	82
5	88
6	92

1. Calculate the **means of X and Y:**

Mean of X ($\bar{X}$) = (2 + 3 + 4 + 5 + 6) / 5 = 4

Mean of Y ($\bar{Y}$) = (65 + 75 + 82 + 88 + 92) / 5 = 80.4

2. Calculate the **differences** between each **data** point and the **means** of X and Y:

X (Hours Studied)	Y (Test Scores)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
2	65	-2	-15.4	30.8
3	75	-1	-5.4	5.4
4	82	0	1.6	0
5	88	1	7.6	7.6
6	92	2	11.6	23.2

3. Calculate the **sum of the products** of the differences ($\Sigma[(X - \bar{X})(Y - \bar{Y})]$):

$\Sigma[(X - \bar{X})(Y - \bar{Y})] = 30.8 + 5.4 + 0 + 7.6 + 23.2 = 67$

4. Calculate the **sum of the squared differences** between X and $\bar{X}$ ($\Sigma[(X - \bar{X})^2]$):

$\Sigma[(X - \bar{X})^2] = (-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10$

5. Calculate the slope (b) using the formula:

$b = \Sigma[(X - \bar{X})(Y - \bar{Y})] / \Sigma[(X - \bar{X})^2$

$= 67 / 10 = 6.7$

6. Calculate the **intercept (a)** using the formula:

a = Mean Y- b* Mean X

$= 80.4 - 6.7 * 4 = 80.4 - 26.8 = 53.6$

So, the **linear regression** equation for this dataset is:

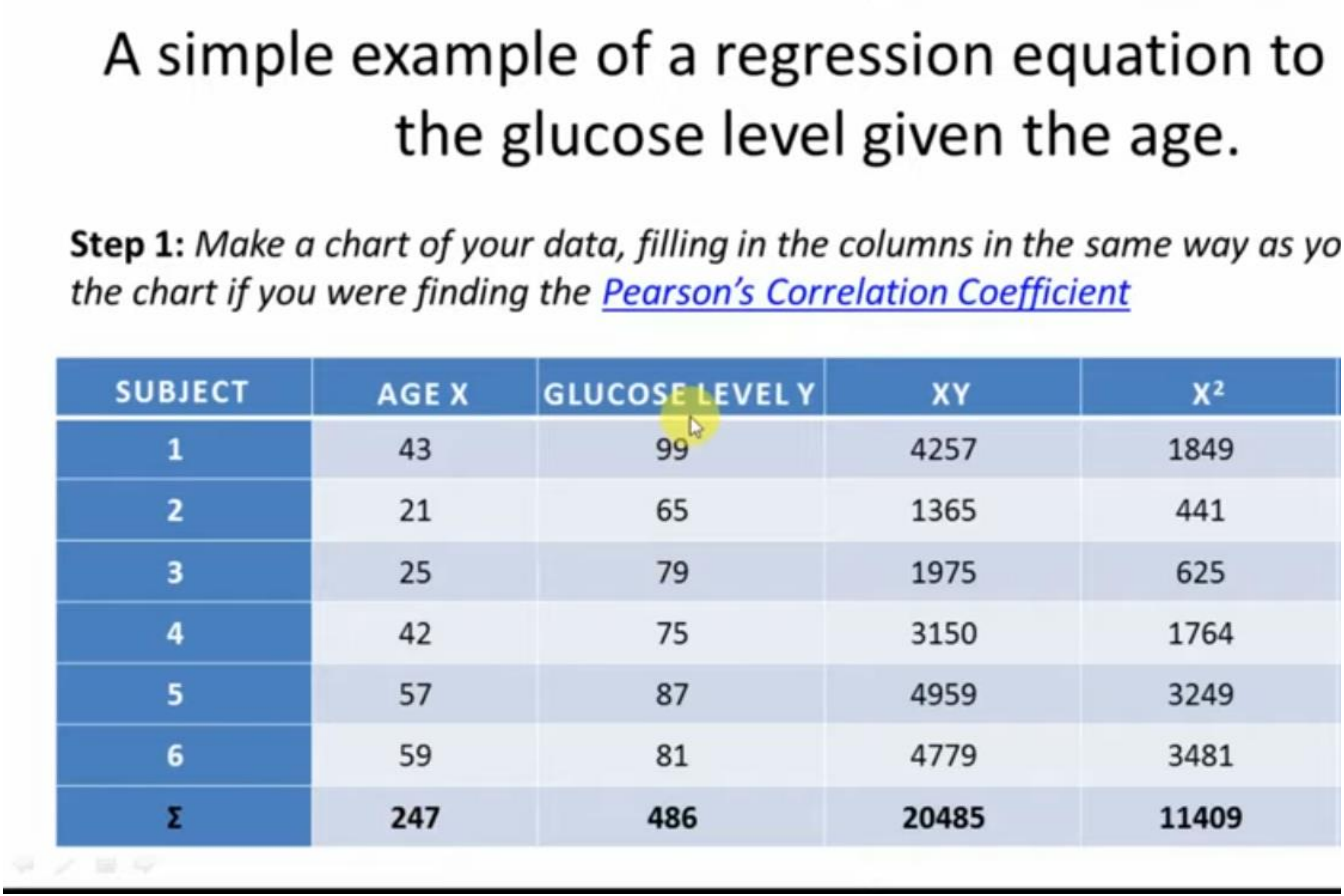
Test Score (Y) = 53.6 + 6.7 * Hours Studied (X)

In this case, the **slope (b)** tells that, on average, for each additional hour studied, the test score is expected to increase by approximately 6.7 points, and the **intercept (a)** is the predicted test score when a student studies for **0 hours.**

With this equation, you can make **predictions** about a student's test score based on the number of hours they study.

For example, if a student studies for **7 hours,** you can predict their test score as:
Test Score = 53.6 + 6.7* 7 = 100.5
Interpretation: according to the regression model, a student who studies for 7 hours is predicted to score around 100.5 on the test.

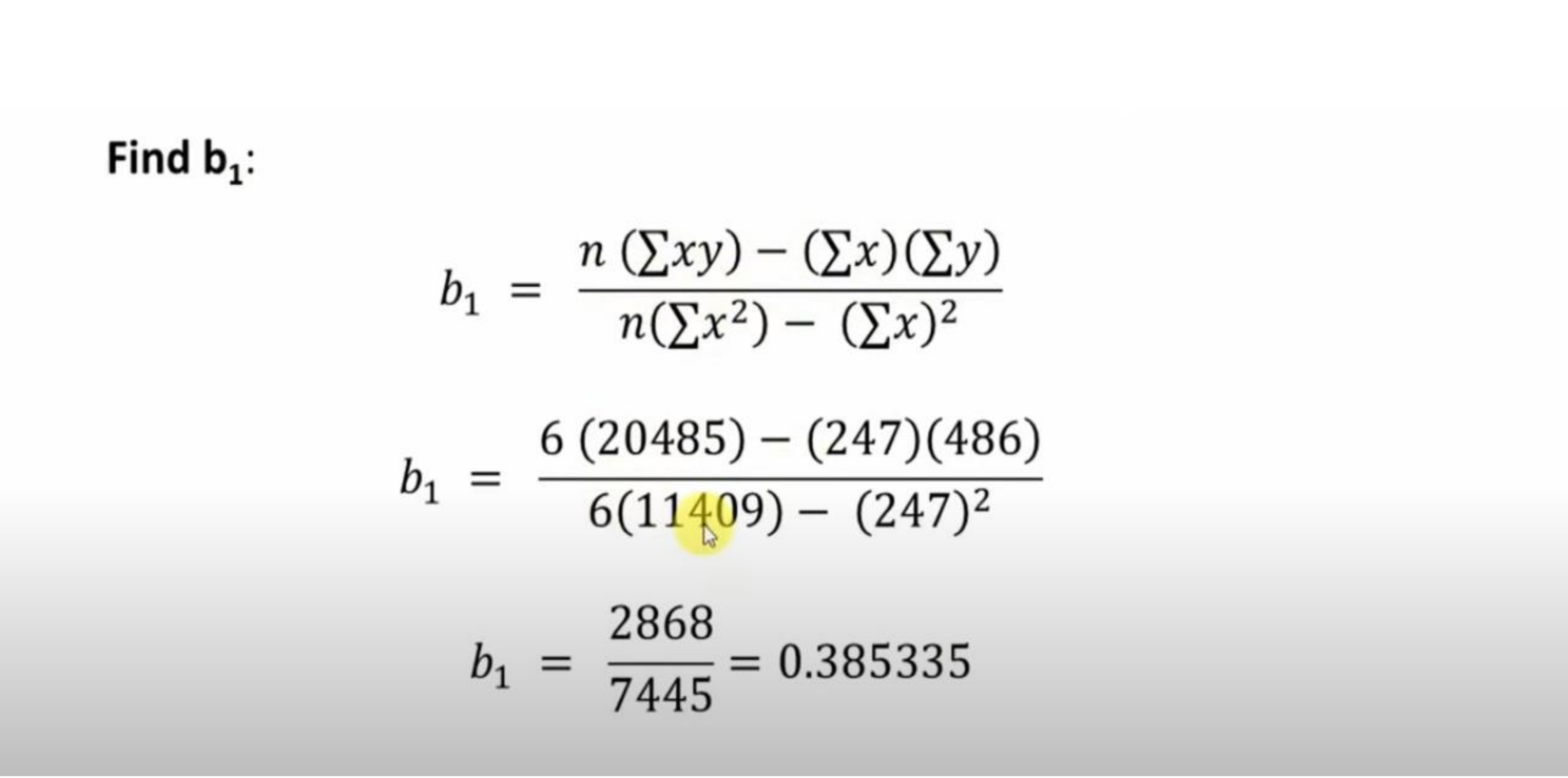
Thus Simple linear regression helps quantify the relationship between variables and make predictions or understand how one variable influences the other.



A simple example of a regression equation to the glucose level given the age.

Step 1: *Make a chart of your data, filling in the columns in the same way as yo the chart if you were finding the* *Pearson's Correlation Coefficient*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2
1	43	99	4257	1849
2	21	65	1365	441
3	25	79	1975	625
4	42	75	3150	1764
5	57	87	4959	3249
6	59	81	4779	3481
Σ	**247**	**486**	**20485**	**11409**



Find b$_1$:

$$b_1 = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{6\,(20485) - (247)(486)}{6(11409) - (247)^2}$$

$$b_1 = \frac{2868}{7445} = 0.385335$$

A simple example of a regression equation to predict the glucose level given the age.

- **Step 3:** *Insert the values into the equation.*

$$y' = b_0 + b_1 x$$

$$\boldsymbol{y' = 65.14 + 0.385225x}$$

- **Step 4:** *Prediction – the value of y for the given value of x = 55*

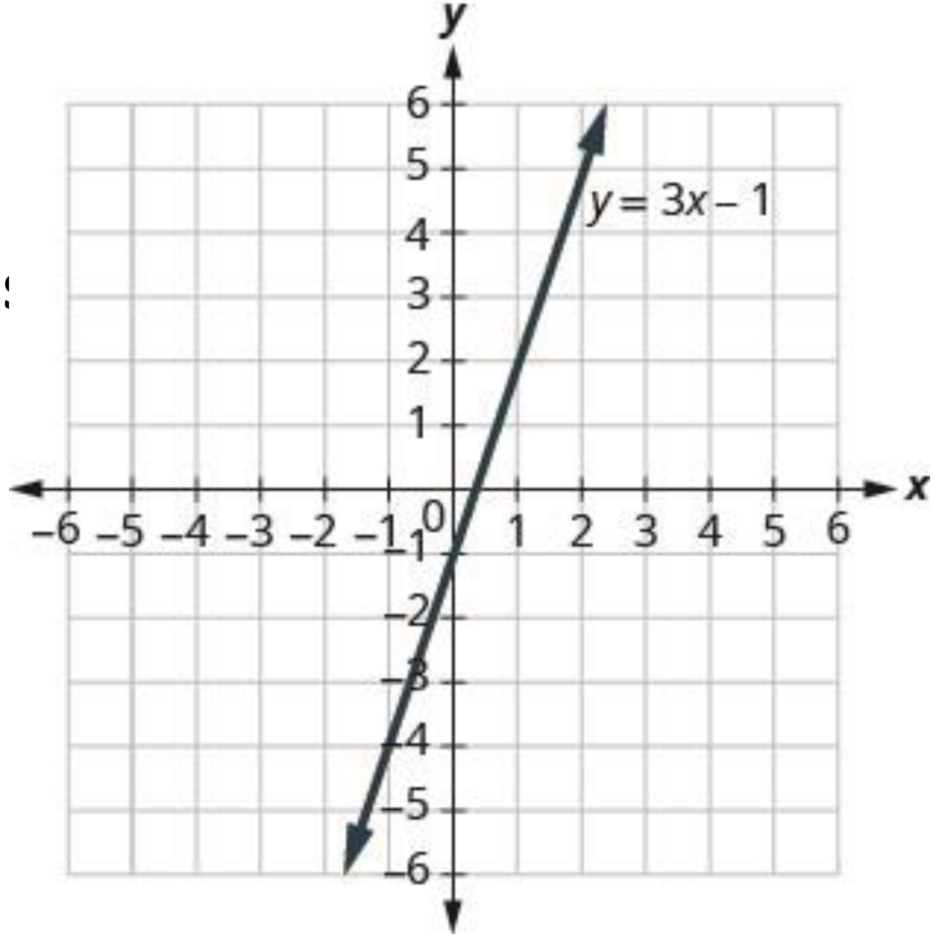
$$\boldsymbol{y' = 65.14 + (0.385225 * 55)}$$

$$\boldsymbol{y' = 86.327}$$

Hence, the glucose level for the given age 55 is 86.327

What does the value a and b mean?

- Suppose if Y= -1+ 3(X), This equation tells us that
- Since **a is negative**, the **line crosses the Y-axis below** the origin. It also describe the value when X=0
- Since the **slope (b) is positive**, the line **extends** from the **lower left-hand corner** to the **upper right hand corner** of the graph.
- It also states that, for each unit increase in X, Y increases by an amount equal to 3.



Practices

- Interpret the following data:
  1. $Y= -3+ 2x$
  2. $Y= 3+0.5\ x$
  3. $Y= 10-0.75X$
  4. $Y= -10-0.75X$
- The following score represent a nurse assessment (X) and a physicians assessment (Y) of the condition of 10 patients at a time of admission to a trauma cancer:

X: 18 13 18 15 10 12 8 4 7 3

Y: 23 20 18 16 14 11 10 7 6 4

Question: Indicate which one you think best fit the data

(a) $y= 8+0.5x$; (b) $y= - 10 + 2x$; (c) $y= 1+1x$

Practice Problems

Q1: Calculate the regression equation of X on Y using method of least squares: $X = 0.5 + 0.5Y$

X	1	2	3	4	5
Y	2	5	3	8	7

Q2: Given the following data:

$N = 8, \Sigma X = 21, \Sigma X^2 = 99, \Sigma Y = 4, \Sigma Y^2 = 68, \Sigma XY = 36$

Using the values, find:

○ Regression Equation of Y on X	$Y = -1.025 + 0.581X$
○ Regression Equation of X on Y	$X = 2.432 + 0.386Y$
○ Value of Y when $X = 10$	$Y = 4.785$
○ Value of X when $Y = 2.5$	$X = 3.397$

Regression Equations Using Regression Coefficients (Using Actual Values)

- Regression Equation of Y on X
  - $Y - \bar{Y} = b_{yx}\,(X - \bar{X})$ where $b_{yx} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$
- Regression Equation of X on Y
  - $X - \bar{X} = b_{xy}\,(Y - \bar{Y})$ where $b_{xy} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2}$

Q3: Calculate the regression equation of Y on X & X on Y

$Y = 1.3X + 1.1$, $X = 0.5 + 0.5Y$

X	1	2	3	4	5
Y	2	5	3	8	7

Regression Equations Using Regression Coefficients (Using Deviations from Actual Values)

- Regression Equation of Y on X
  - $\mathrm{Y} - \bar{Y} = \mathrm{b_{yx}}\ (\mathrm{X} - \bar{X})$ where $\mathrm{b_{yx}} = \frac{\Sigma xy}{\Sigma x^2}$
- Regression Equation of X on Y
  - $\mathrm{X} - \bar{X} = \mathrm{b_{xy}}\ (\mathrm{Y} - \bar{Y})$ where $\mathrm{b_{xy}} = \frac{\Sigma xy}{\Sigma y^2}$

Q4: Calculate the regression equation of Y on X & X on Y using method of least squares: $\mathrm{Y} = 0.26\mathrm{X} + 3.2$, $\mathrm{X} = 4.75 + 0.45\mathrm{Y}$

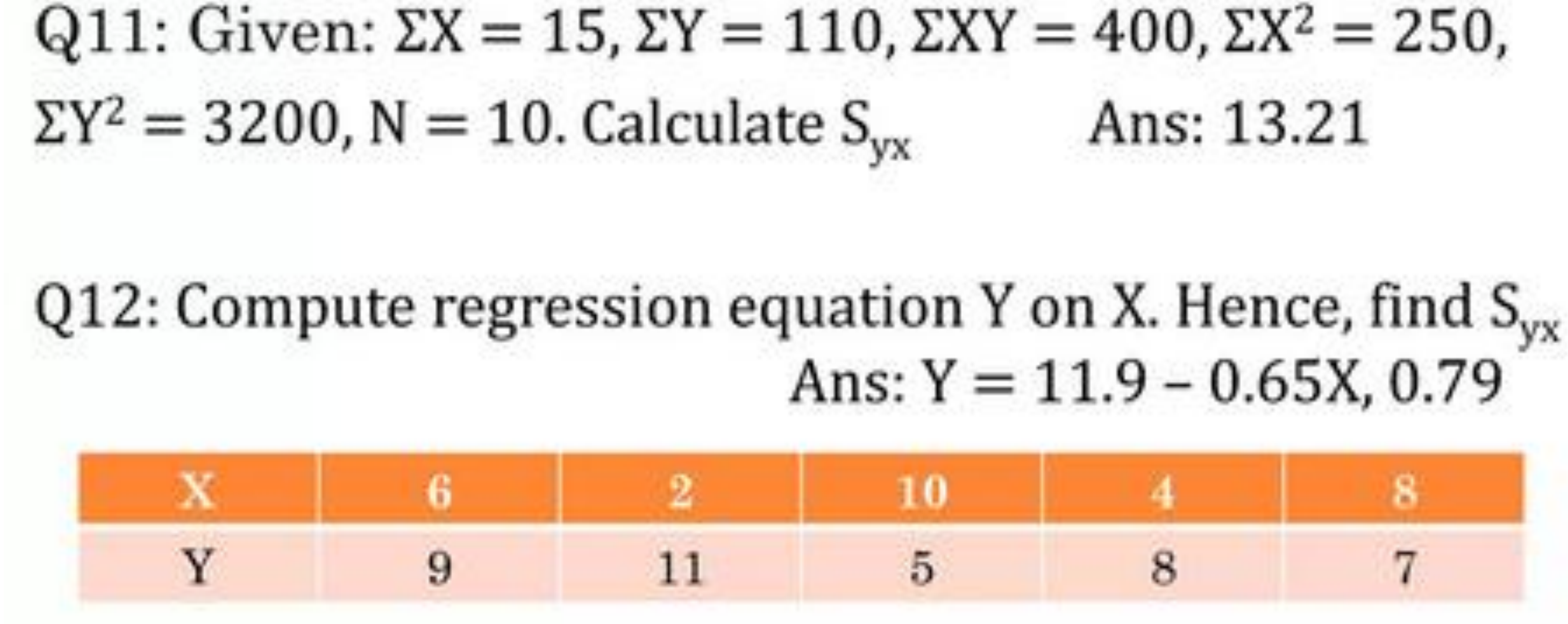
X	2	4	6	8	10	12
Y	4	2	5	10	3	6

Standard Error of Estimate or Regression

- The standard error of the regression (S), also known as the standard error of the estimate, **represents the average distance that the observed values fall from the regression line**.
- A **small SE** is an indication that the **sample mean is a more accurate** reflection of the **actual population mean**. A **larger sample size** will normally **result** in a **smaller SE** (while SD is not directly affected by sample size).
- For two regression lines, there are two standard error of estimates

  1. SE of estimates of Y on X (S_{yx})
  2. SE of estimates of X on Y (S_{xy})

- Formula: $\mathbf{S_{yx} = \surd((\Sigma Y^2 - a\Sigma Y - b\Sigma XY)/N})$;
- here a and b are to be obtained from normal equation.
- Solve this:



Q11: Given: $\Sigma X = 15$, $\Sigma Y = 110$, $\Sigma XY = 400$, $\Sigma X^2 = 250$, $\Sigma Y^2 = 3200$, $N = 10$. Calculate S_{yx} Ans: 13.21

Q12: Compute regression equation Y on X. Hence, find S_{yx}

Ans: $Y = 11.9 – 0.65X$, 0.79

X	6	2	10	4	8
Y	9	11	5	8	7

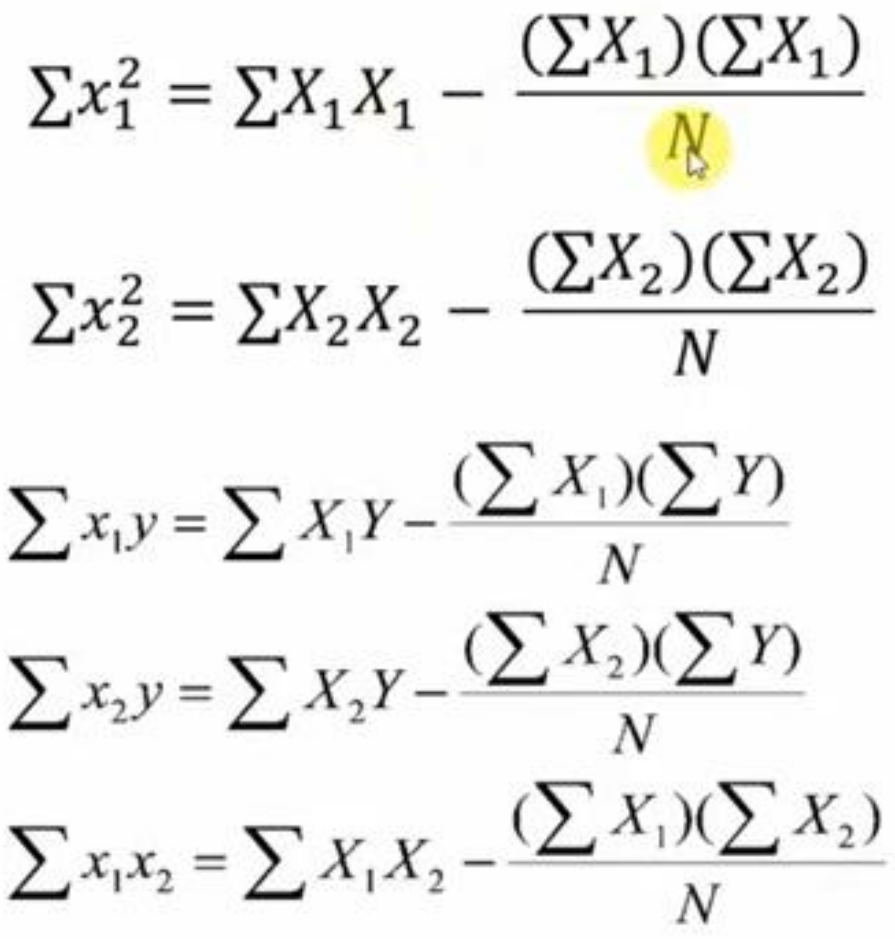
Multiple Linear Regression Analysis

- Here the study variable(outcome) depends on **more than one explanatory or independent variables**, called a multiple linear regression model.
- Let Y denotes the dependent (or study) variable that is linearly related to K independent (or explanatory) variables X_1, X_2, ,..., X_k through the parameters β_1, β_2, ,..., β_k and we write

$$\mathbf{Y = X_1\beta_1 + X_2\beta_2 + \ldots\ldots X_k\beta_k + \varepsilon}$$

- The parameters **β_1, β_2, ,..., β_k are the regression coefficients associated** with **X_1, X_2, ,..., X_k** respectively and **ε is the random error** component **reflecting the difference between the observed and fitted linear relationship.**

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
∑	19.5	20	24	90	148	99	95.8	45.6



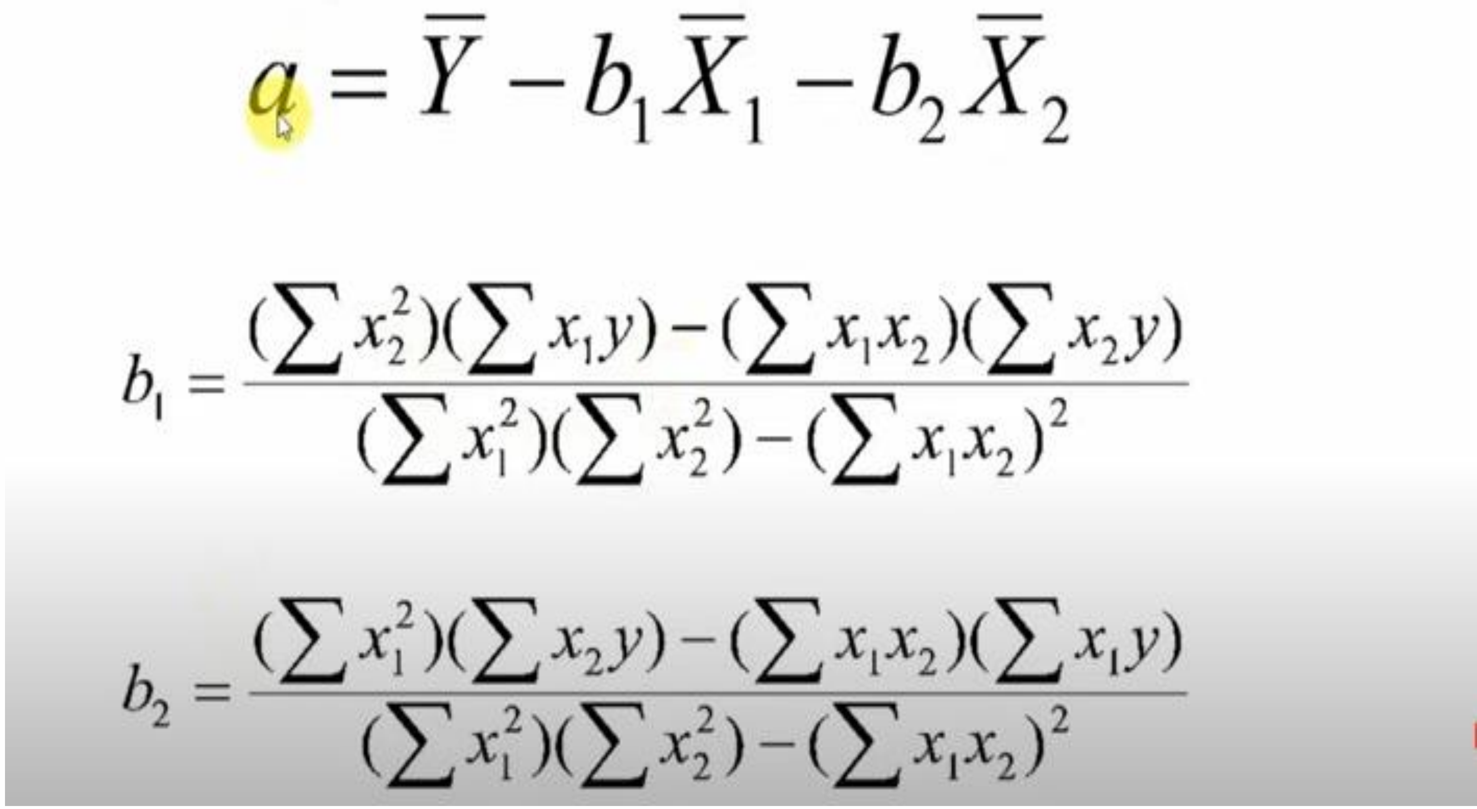
$$\sum x_1^2 = \sum X_1X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\sum x_2^2 = \sum X_2X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\sum x_1y = \sum X_1Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2y = \sum X_2Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1x_2 = \sum X_1X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$



$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2y) - (\sum x_1x_2)(\sum x_1y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{32.8 \ast 17.8 \; - 3 \ast (-48)}{10 \ast 32.8 \; - 3 \ast 3} = 2.28$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{10 \ast (-48) \; - 3 \ast 17.8}{10 \ast 32.8 \; - 3 \ast 3} = -1.67$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 = \frac{19.5}{5} - \frac{2.28 \; \ast 20}{5} - \frac{-1.67 \ast 24}{5} = 2.796$$

Final Regression equation or Model is:

$$\boldsymbol{Y = 2.796 + 2.28\,x_1 - 1.67x_2}$$

$$Now\ given \quad x_1 = 3 \quad and \quad x_2 = 2 \qquad Y = ?$$

$$\begin{aligned} Y &= 2.796 + 2.28 * 3 - 1.67 * 2 \\ &= \mathbf{6.296} \end{aligned}$$

Example2:
Step 1: Calculate Regression Sums.

make the following regression sum calculations:

- $\Sigma x_1^2 = \Sigma X_1^2 – (\Sigma X_1)^2$ / n = 38,767 – $(555)^2$ / 8 = **263.875**

- $\Sigma x_2^2 = \Sigma X_2^2 – (\Sigma X_2)^2$ / n = 2,823 – $(145)^2$ / 8 = **194.875**

- $\Sigma x_1y = \Sigma X_1y – (\Sigma X_1\Sigma y)$ / n = 101,895 – (555*1,452) / 8 = **1,162.5**

- $\Sigma x_2y = \Sigma X_2y – (\Sigma X_2\Sigma y)$ / n = 25,364 – (145*1,452) / 8 = **-953.5**

- $\Sigma x_1x_2 = \Sigma X_1X_2 – (\Sigma X_1\Sigma X_2)$ / n = 9,859 – (555*145) / 8 = **-200.375**

	y	X_1	X_2
	140	60	22
	155	62	25
	159	67	24
	179	70	20
	192	71	15
	200	72	14
	212	75	14
	215	78	11
Mean	181.5	69.375	18.125
Sum	1452	555	145

	X_1^2	X_2^2	X_1y	X_2y	X_1X_2
	3600	484	8400	3080	1320
	3844	625	9610	3875	1550
	4489	576	10653	3816	1608
	4900	400	12530	3580	1400
	5041	225	13632	2880	1065
	5184	196	14400	2800	1008
	5625	196	15900	2968	1050
	6084	121	16770	2365	858
Sum	38767	2823	101895	25364	9859

Reg Sums	263.875	194.875	1162.5	-953.5	-200.375

Step 2: Calculate b_0, b_1, and b_2.

The formula to calculate b_1 is: $[(\Sigma x_2^2)(\Sigma x_1 y) - (\Sigma x_1 x_2)(\Sigma x_2 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_1} = [(194.875)(1162.5) - (-200.375)(-953.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{3.148}$

The formula to calculate b_2 is: $[(\Sigma x_1^2)(\Sigma x_2 y) - (\Sigma x_1 x_2)(\Sigma x_1 y)] / [(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2]$

Thus, $\mathbf{b_2} = [(263.875)(-953.5) - (-200.375)(1152.5)] / [(263.875)(194.875) - (-200.375)^2] = \mathbf{-1.656}$

The formula to calculate b_0 is: $y - b_1 X_1 - b_2 X_2$

Thus, $\mathbf{b_0} = 181.5 - 3.148(69.375) - (-1.656)(18.125) = \mathbf{-6.867}$

Step 3: Place b_0, b_1, and b_2 in the estimated linear regression equation.

The estimated linear regression equation is: $\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$

In our example, it is $\mathbf{\hat{y} = -6.867 + 3.148x_1 - 1.656x_2}$

How to Interpret a Multiple Linear Regression Equation

Here is how to interpret this estimated linear regression equation:

$$\hat{y} = -6.867 + 3.148x_1 - 1.656x_2$$

b_0 = -6.867. When both predictor variables are equal to zero, the mean value for y is -6.867.

b_1 = 3.148. A one unit increase in x_1 is associated with a 3.148 unit increase in y, on average, assuming x_2 is held constant.

b_2 = -1.656. A one unit increase in x_2 is associated with a 1.656 unit decrease in y, on average, assuming x_1 is held constant.

What Is R-Squared or Coefficient of regression?

- R-squared (R^2) is a statistical measure that represents the **proportion of the variance for a dependent variable** that's explained by an independent variable in a regression model.
- Whereas **correlation** explains the **strength of the relationship** between an independent and a dependent variable, R-squared explains the **extent** to which the **variance of one variable explains the variance of the second variable**. So, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.
- R-squared values range from 0 to 1 and are commonly stated as percentages from 0% to 100%. An R-squared of 100% means that all of the movements of a security (or another dependent variable) are completely explained by movements in the index (or whatever independent variable you are interested in).

Adjusted **R-Squared**

- R-squared only works as **intended** in a **simple linear** regression model with one explanatory variable. With a **multiple** regression made up of several independent variables, the **R-squared must be adjusted.**
- The adjusted R-squared **compares** the **descriptive power** of regression models that include diverse numbers of predictors.
- Every **predictor** added to a model **increases R-squared** and never decreases it. Thus, a model with **more terms** may seem to have a **better fit** just for the fact that it has more terms.
- It only **increases** if the **new term enhances** the model **above** what would be **obtained by probability** and **decreases** when a predictor **enhances** the model **less than what is predicted** by chance.

Differences between correlation and regression

Characteristic	Correlation	Regression
Degree and nature of relationship	Correlation is a measure of degree of relationship between X and Y	Regression studies the nature of relationship between the variables so that one may be able to predict the value of one variable on the basis of another
Cause and effect relationship	Doesn't always assume cause and effect relationship between two variables	Express the cause and effect relationship between two variables. The independent is the cause and dependent is the effect
Prediction	Doesn't help in making prediction	Enable us to make prediction using regression line
Symmetric	Correlation coefficient are symmetrical	Regression coefficient are not symmetrical
Origin and scale	It is independent of the change of origin and scale	RC is independent of change of origin but not of scale.

Class Work

- **What is the differences between simple linear regression and multiple linear regression?**
- **Try to find the relation between the GPA of a class of students, the number of hours of study, and the student's height.**

GPA (Y)	Hr of Study (X1)	Height (inch) (X2)
3.02	2	57
3.72	7	53
3.55	5	60
3.45	6	55