Correlation measures the strength and direction of linear association between the variables, but is does not tell us the effect of one variable to another. In this case, regression analysis is used which objective is to model the association with a line and give the equation.

1. Regression: Regression refers to the cause-and-effect relationship between two or more variables, where affected variables are dependent variables and causal variables are independent variables. This relationship is mathematically expressed, highlighting the impact of independent variables on dependent variables.

2. Correlation vs Regression:

Correlation	Regression
1. Correlation is a statistical measure that determines the association between two variables.	1. Regression describe how to numerically relate an independent variable to the dependent variable.
2. There is no dependent variable and independent variable.	2. Must be one dependent variable and one independent variable.
3. To represent the linear relationship between variables.	3. To represent the cause-and-effect relationship between variables.

3. Types of regression model: On the basis of variables, there are two types of regression model,

a) ***Simple linear regression model:*** Simple Linear Regression Model is a statistical method used to study the relationship between two continuous/numerical variables, typically denoted as X (independent variable) and Y (dependent variable). The goal of this model is to establish a linear relationship between these variables, represented by the equation:

$$Y_i = \alpha + \beta X_i + \epsilon_i \quad ; i = 1,2,3, \dots, n$$

Here,

$Y = Dependent\ variable$

$X = Independent\ variable$

$\alpha = Intercept\ coefficient$

$\beta = Slope\ coefficieint$

$\epsilon = Random\ error\ term$

b) ***Multiple linear regression model:*** Multiple Linear Regression Model is a statistical method used to study the relationship between one dependent continuous variable and more than two independent continuous variables, typically denoted as X (independent variable) and Y (dependent variable). The goal of this model is to establish a linear relationship between these variables, represented by the equation:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \quad ; i = 1,2,3, \dots, n$$

Here,

$Y = Dependent\ variable$

$X = Independent\ variable$

$\alpha = Intercept\ coefficient$

$\beta_1, \dots, \beta_k = Slope\ coefficieints$

$\epsilon = Random\ error\ term$

4. Basic assumptions of regression model:

a) The values of the independent variable X are fixed in advance.

b) The values of the dependent variable Y depend on the values of X.

c) Random error is independently normally distributed with mean zero and variance σ^2

5. Estimating regression parameters ($\boldsymbol{\alpha}$ *and* $\boldsymbol{\beta}$): Let the simple linear regression model of dependent variable Y on independent variable X be,

$$Y = \alpha + \beta X + \epsilon$$

Where, α and β be known regression parameter. So, we need to estimate these parameters. We may estimate these parameters by using Ordinary Least Square (OLS) estimation method/ Method of Least Square. By applying OLS method, the estimated regression parameters can be written as,

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

and,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = \frac{\sum Y_i}{n} - \hat{\beta}\frac{\sum X_i}{n}$$

Then, the estimated regression model/fitted regression model can be written as,

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X$$

We can calculate the value of random error term:

$$\boldsymbol{\epsilon = Y - \hat{Y}}$$

Math: Find the fitted regression line for sales revenue on advertising cost. Also, estimate sales revenue when advertising cost is $9

Advertising cost ($ million)	Sales revenue ($ million)
2	7
1	3
3	8
4	10

Solution:

First, we construct a table for calculation of regression parameters.

Advertising cost (x_i)	Sales revenue (y_i)	x_i^2	$x_i \times y_i$
2	7	4	14
1	3	1	3
3	8	9	24
4	10	16	40
$\sum x_i = 10$	$\sum y_i = 28$	$\sum x_i^2 = 30$	$\sum x_i y_i = 81$

Now, the estimated slope of the regression line is,

$$\hat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{(4 \times 81) - (10 \times 28)}{(4 \times 30) - (10)^2} = 2.2$$

Interpretation of β: $\hat{\beta} = 2.2$ means that an increase of $1 million in advertising cost, the average sales revenue will increase $2.2 million.

Now, the estimated intercept of the regression line is,

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 7 - (2.2 \times 2.5) = 1.5$$

Interpretation of α: $\hat{\alpha} = 1.5$ means that, if there is no advertising cost then average sales revenue would be $1.5 million.

So, the fitted regression equation is, $\widehat{Y}_i = 1.5 + 2.2X_i$.

$If\ X = 9$: $\widehat{Y}_i = 1.5 + 2.2 \times 9 = 21.3$

6. Coefficient of determination: The coefficient of determination tells the percent of the variation in the dependent variable that is explained (determined) by the model and the explanatory variable. It is denoted by R^2.

$$R^2 = 1 - \frac{SSE}{SST}$$

Where,

$SSE = \sum (Y_i - \hat{Y})^2$

$SST = \sum (Y_i - \bar{Y})^2$

Advertising cost (x_i)	**Sales revenue** (y_i)	$\hat{y} = 1.5 + 2.2x_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
2	7	5.9	1.21	0
1	3	3.7	0.49	16
3	8	8.1	0.01	1
4	10	10.3	0.09	9
$\sum x_i = 10$	$\sum y_i = 28$		$SSE = 1.8$	$SST = 26$

$$R^2 = 1 - \frac{SSE}{SST}$$

$$R^2 = 1 - \frac{1.8}{26}$$

$$R^2 = 0.93$$

Interpretation of R^2: For example, $R^2 = 0.927$ *or* 92.7%. It indicates that, almost 93% of the variability of the dependent variables explained by the independent variables.

Error calculation formula, $\epsilon_i = (y_i - \hat{y}_i)$

Advertising cost (x_i)	**Sales revenue** (y_i)	$\hat{y} = 1.5 + 2.2x_i$	$\epsilon_i = (y_i - \hat{y}_i)$
2	7	5.9	1.1
1	3	3.7	-0.7
3	8	8.1	-0.1
4	10	10.3	-0.3

Extra:

1. The following data relate to the volume of investment and the corresponding amount of profit per year.

Investment ($ million)	5	10	15	20	25
Profit ($ million)	3	4	8	12	18

a) Is there any relationship between investment and profit? Explain mathematically. $(Ans: r = 0.97)$

b) Estimate linear regression line of profit on investment. [$\boldsymbol{Tips}$: $\boldsymbol{Y\ on\ X}$]

c) Compute the expected amount of profit if the volume of investment is 32$ (million). $(Ans: Expected\ Profit = 21.92\$\ Million)$

d) Compute the volume of investment if the expected amount of profit is 20$ (million). $(Ans: Volume\ of\ investment = 29.47\$\ Million)$

2.

Given below are the advertisement expenditure (in thousand taka) and sales volume (in thousand taka) of a company:

Advertisement expenditure	Sales volume
72	400
60	392
84	442
108	421
142	570
135	450

(i) Compute coefficient of correlation between advertisement expenditure and sale.

(ii) Find regression equation of sale on advertisement expenditure.

(iii) Find expected amount of sale if the advertisement expenditure is 150 thousand taka.

(iv) Find advertisement expenditure if the expected amount of sale is 600 thousand taka.

3.

Amount of fertilizer used (in kg) on same size of plots and amount of corn yield (in quintal) are given in the following table.

Fertilizer	Corn yield
3	10
6	15
9	30
12	35
15	25
18	30
21	50
24	45

(i) Plot the dependent variable against the independent variable.
(ii) Find the least squares line for this data.
(iii) What is the y-intercept? If 30 kg of fertilizer were used, what would be expected amount of corn yield?

4.

From the following data, find (i) the regression line of y on x and (ii) correlation coefficient of x and y.

$\Sigma x = 56, \Sigma y = 40, \Sigma x^2 = 524, \Sigma y^2 = 256, \Sigma xy = 364, n = 8.$