Linear Regression Analysis

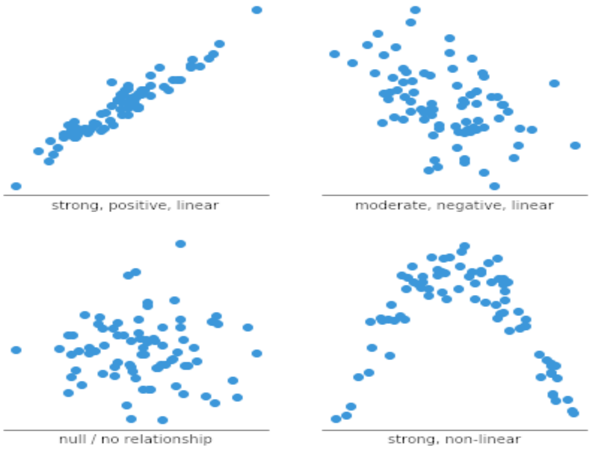
Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

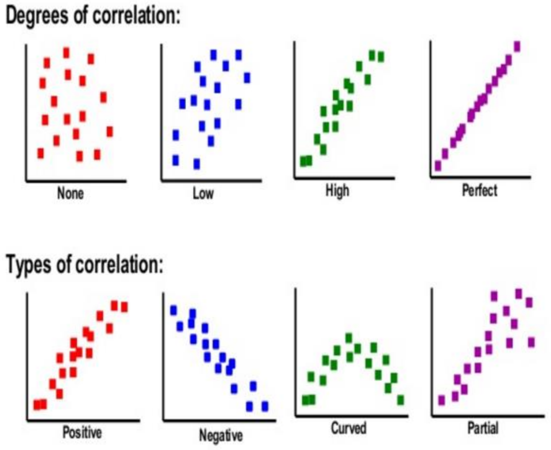
There are generally two types of variables in regression analysis. The variable whose value is influenced or is to be predicted is called dependent/regressed/responsed variable and the variable which influences the value or is used for prediction is called independent/regressor/predictor/ explanatory variable.

Scatter Diagram:

This is the simplest diagrammatic representation of bivariate data. For a bivariate distribution $x_i|y_i$ $(i = 1, 2, \ldots.., n)$, if the values of the variables x and y are plotted along the x-axis and y-axis respectively in the x-y plane, the diagram of dots so obtained is known as the scatter diagram. From this diagram, one can have an idea whether the variables are correlated or not. If the points are very dense, a fairly good amount of correlation between the variables is expected whereas if the points are widely scattered, a poor correlation is expected. However, this method is not suitable if the number of observations is large.

If the variables are related in a bivariate distribution, then the points are found in the scatter diagram clustering round some curve, called the curve of regression and there is said to be curvilinear regression between the variables. If the curve is a straight line, then it is called the line of regression and in this case, there is said to be linear regression between the variables.





Linear Regression:

The line of regression gives the best estimate to the value of one variable for any specific value of the other variable. Thus it is called the line of 'best fit' and can be derived from the principle of least squares.

Let in a bivariate distribution $x_i | y_i \ (i = 1, 2, \dots.., n)$, y is dependent variable and x is independent variable. Then the line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) - - - - (1)$$

where $\bar{x}$ and $\bar{y}$ are the means of the two distributions; $b_{yx} = \frac{\text{cov}(x,y)}{\sigma_x{}^2} = r\frac{\sigma_y}{\sigma_x}$ is called the coefficient of regression of y on x; r is Karl Pearson's correlation coefficient; σ_x and σ_y are the standard deviations of the two distributions. It can be easily seen that b_{yx} gives the estimate of slope of the regression line of y on x. This line is used to estimate or predict the value of y for any given value of x. This estimate will be best in the sense that it will have the minimum possible errors as defined by the principle of least squares.

Let us now assume that x is dependent variable and y is independent variable. Then the line of regression of x on y will be given by

$$x - \bar{x} = b_{xy}(y - \bar{y}) - - - - (2)$$

In this case, $b_{xy} = \frac{\text{cov}(x,y)}{\sigma_y{}^2} = r\frac{\sigma_x}{\sigma_y}$ is called the regression coefficient of x on y and gives the estimate of slope of the regression line of x on y. This line is used to estimate or predict the value of x for any given value of y. This estimate will be best in the sense that it will have the minimum possible errors as defined by the principle of least squares.

Notes:

(i) The point of intersection of the two regression lines (1) and (2) is given by $(\bar{x}, \bar{y})$.

(ii) In the case of perfect correlation (i.e, for $r = \pm 1$), both the lines of regression coincide in a single line

$$\frac{y-\bar{y}}{\sigma_y} = \pm\left(\frac{x-\bar{x}}{\sigma_x}\right)$$

(iii) We have $b_{xy} \times b_{yx} = r^2 \rightarrow r = \pm\sqrt{b_{xy} \times b_{yx}}$

(iv) We know that: $r = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y}$, $b_{xy} = \frac{\text{cov}(x,y)}{{\sigma_y}^2}$ and $b_{yx} = \frac{\text{cov}(x,y)}{{\sigma_x}^2}$

Hence the sign of r will be same as that of b_{xy} and b_{yx}. This means all three of them will have the same sign as that of cov (x, y).

(v) If one of the regression coefficient is greater than unity, then the other must be less than unity.

(vi) Regression coefficients are independent of the change of origin but not of scale

(vii) If θ is the angle between the two lines of regression, then it is given as

$$\theta = \tan^{-1}\left[\frac{1-r^2}{|r|}\left(\frac{\sigma_x\sigma_y}{{\sigma_x}^2+{\sigma_y}^2}\right)\right]$$

- If $r = 0$, then $\theta = \frac{\pi}{2}$ → This means when two variables are uncorrelated, then the two lines of regression become perpendicular to each other

- If $r = \pm 1$, then $\theta = 0,\ \pi$ → In this case, the two lines either coincide or they are parallel to each other. But we have already known the fact that both the lines pass through the point $(\bar{x}, \bar{y})$, hence they cannot be parallel. Therefore they must coincide. [*refer to note (ii)]

Problems:

Ex.1. (i) Find the regression equation of y on x from the following data: $b_{yx} = 1.5$, $\bar{x} = 5$ and $\bar{y} = 7$

(ii) If $\sigma_x = 10$, $\sigma_y = 12$, $b_{xy} = -0.8$, find r

(iii) The regression coefficients of y on x and x on y are (– 1.2) and (- 0.3) respectively. Find the correlation coefficient.

(iv) The regression equation of y on x is 3x – 4y + 61 = 0. If the mean of y is 52, find the mean of x.

(v) The regression equation of y on x is 3x – 5y = 13 and the regression equation of x on y is 2x – y = 7. Estimate the value of x when y = 10.

Solution: (i) The regression equation of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or}, y - 7 = 1.5(x - 5) \ \ \text{or}, y = 1.5x - 0.5$$

(ii) From the definition, $b_{xy} = r\frac{\sigma_x}{\sigma_y} \to r = \frac{b_{xy}\times\sigma_y}{\sigma_x} = -\frac{0.8\times 12}{10} = -0.96$

(iii) Here we have $b_{yx} = -1.2$ and $b_{xy} = -0.3 \to r^2 = 0.36 \to r = -0.6$

(iv) We have $\bar{y} = 52$, we also know that the regression lines pass through the point $(\bar{x}, \bar{y})$. Therefore the line 3x – 4y + 61 = 0 must pass through the point $(\bar{x}, 52)$, i.e,

$$3\bar{x} - 4\times 52 + 61 = 0 \ \ \text{or}, \ \ \bar{x} = 49$$

(v) Since we need to estimate x when y = 10, therefore x is dependent variable and y is independent variable. So we need to use the regression equation of x on y. From 2x – y = 7, when y = 10, we get x = 8.5.

Ans.

Ex.2. Find the most likely price in Mumbai corresponding to the price of Rs. 70 at Ahmedabad from the following:

	Ahmedabad	Mumbai
Average price	65	67
Standard Deviation	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8.

Solution: Let the price in Ahmedabad and Mumbai be x and y respectively. Then we have $\bar{x} = 65, \bar{y} = 67, \sigma_x = 2.5, \sigma_y = 3.5, r = 0.8$. We need to predict the value of y when x = 70, which means we need the regression equation of y on x. This is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$\text{or}, y - 67 = \frac{0.8 \times 3.5}{2.5}(x - 65)$$

From the above expression, when x = 70, y = 72.6. Hence, the required price in Mumbai will be Rs. 72.6.

Ans.

Ex.3. Marks obtained by 12 students in the college test (x) and the university test (y) are as follows:

x	41	45	50	68	47	77	90	100	80	100	40	43
y	60	63	60	48	85	56	53	91	74	98	65	43

What is your estimate of the marks a student could have obtained in the university test if he obtained 60 in the college test but was ill at the time of the university test?

Solution: Since we need to estimate y for a given value of x = 60, we need the regression equation of y on x given by

$$y-\bar{y}=b_{yx}(x-\bar{x})----(1)$$

where the regression coefficient b_{yx} is given by

$$b_{yx}=\frac{\operatorname{cov}(x,y)}{{\sigma_x}^2}$$

Let us make the table for the calculation of the regression line:

x	41	45	50	68	47	77	90	100	80	100	40	43	781
y	60	63	60	48	85	56	53	91	74	98	65	43	796
x^2	1681	2025	2500	4624	2209	5929	8100	10000	6400	10000	1600	1849	56917
xy	2460	2835	3000	3264	3995	4312	4770	9100	5920	9800	2600	1849	53905

The means of the two distributions are $\bar{x}=\frac{781}{12}=65.08,\ \bar{y}=\frac{796}{12}=66.33$

The variance ${\sigma_x}^2$ is given by

$${\sigma_x}^2=\frac{1}{n}\sum_{i=1}^{n}{x_i}^2-\bar{x}^2=\frac{56917}{12}-65.08^2=4743.08-42\;35.41=507.67$$

The covariance between x and y is given

$$\text{cov}(x,y) = \frac{\sum xy}{n} - \bar{x}\,\bar{y} = \frac{53905}{12} - 65.08 \times 66.33 = 4492.08 - 4316.76$$

$$= 175.32$$

Therefore,

$$b_{yx} = \frac{\text{cov}(x,y)}{{\sigma_x}^2} = \frac{175.32}{507.57} = 0.345$$

Then from the equation (1), the required regression line will be

$$y - 66.33 = 0.345(x - 65.08)$$

$$\text{or, } y = 0.345x + 43.88$$

Hence, for x = 60, y = 0.345×60 + 43.88 = 64.58 ≈ 65.

Ans.