

## Linear Regression

Regression can be defined as a method to estimate the value of one variable when that of other is known, when the variables are correlated. Regression analysis is a mathematical measure of average relationship between two or more correlated values.

### Linear regression

1. Line of regression of y on x is:

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

where regression coefficient of y on x is given by

$$b_{yx} = \frac{\text{Cov}(x,y)}{\sigma_x^2} = r \frac{\sigma_y}{\sigma_x}$$

2. Line of regression of x on y is:

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

where regression coefficient of x on y is given by

$$b_{xy} = \frac{\text{Cov}(x,y)}{\sigma_y^2} = r \frac{\sigma_x}{\sigma_y}$$

**Properties:**

1. Lines of regression passes through the point  $(\bar{x}, \bar{y})$
2.  $b_{yx}b_{xy} = r^2$ . [ $r = \sqrt{b_{yx} \cdot b_{xy}}$ ]
3.  $b_{yx}, b_{xy}$  have same sign.

### Least Squares Straight Line

For a given set of  $N$  data points  $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$   
assume that the straight line

$$Y = a_0 + a_1 X = f(X)$$

fits to the data in the least squares sense  
Normalized equation are given by

$$\sum Y_i = N a_0 + a_1 \sum X_i$$

$$\sum X_i Y_i = a_0 \sum X_i + a_1 \sum X_i^2$$

known as "**Normal equations**".

1. The following are the marks in Statistics (X) and Mathematics(Y) of ten students

X	56	55	58	57	56	60	54	59	57	58
Y	68	67	67	65	68	70	66	68	66	70

Calculate the coefficient of correlation and estimate marks in Mathematics of a student who scored 62 marks in Statistics.

**[Ans :  $r = 0.44$  ,  $Y = 69.5$  ]**



2. It is given that the means of  $x$  and  $y$  are 5 and 10. If the line of regression of  $y$  on  $x$  is parallel to the line  $20y = 9x + 40$ , estimate the value of  $y$  at  $x = 30$

**[Ans:  $20y = 9x + 155$  and  $y = 21.25$ ]**



3. Find the two lines of regression from the following data

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

[Ans :  $x = 30.364 + 0.545 y$  and  $y = 23.667 + 0.667 x$ ]





- .
4. In partially destroyed laboratory record of an analysis of correlation data, the following results only are legible- Variance of  $X = 9$ , regression equations are:  $8X - 10Y + 66 = 0$  &  $40X - 18Y = 214$  What was
- (a) the mean of  $X$  and  $Y$
  - (b) the correlation between  $X$  and  $Y$
  - (c) the S.D. of  $Y$



5. You are given the following data

	X	Y
Mean	30.1	47.8
standard deviation	6.2	9.5

6. Obtain the equation of the line of regression of cost on age from the following table giving the age of a car of certain make and the annual maintenance cost.

Age of car(in years)	2	4	6	8
Maintenance (in thousand of Rs.)	5	7	8.5	11

Also find maintenance cost of the car if its age is 9 years

**[Ans :  $y = 3 + 0.975 x$  and  $y = \text{Rs. } 11775$ ]**

a