Regression Analysis:

If the variables X and Y in a bivariate distribution are related we will find that the points in the scatter diagram will cluster around some curve called the Curve of Regression. If the curve is a straight line then it is called the Line of Regression and there is said to be a linear regression between the variables. The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other. In fact, there are two such lines one giving the best possible mean values of Y for each specified value of X and other giving the best possible mean values of X for each specified value of Y. The former is known as the line of regression of Y on X and the latter is known as the line of regression of X on Y. Thus the line of regression is the line of best fit and is obtained using the principles of least squares.

Note: The principle of least squares consists in minimizing the sum of squares of the deviations of the actual values of Y from its estimated values as given by the line of best fit.

Thus, the line of regression of Y on X is given by;

$$(Y - \overline{Y}) = r\frac{\sigma_Y}{\sigma_X}(X - \overline{X}) \longrightarrow \quad (A)$$

and line of regression of X on Y is given by;

$$(X - \overline{X}) = r\frac{\sigma_X}{\sigma_Y}(Y - \overline{Y}) \longrightarrow \quad (B)$$

where 'r' is the sample correlation coefficient and σ_X and σ_Y are the standard deviations of X and Y respectively

We denote the factors:

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$ and is called the regression coefficient of Y on X.

$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$ and is called the regression coefficient of X on Y.

NOTE:

1. Whenever we have to estimate Y for a given value of X, i.e., Y is dependent and X is independent, we then use equation (A) otherwise we use (B).
2. In the case of perfect correlation i.e., $r = \pm 1$, the two lines of regression coincide. Thus we have only one line.
3. The correlation coefficient is obtained as the geometric mean of the two regression coefficients. Thus $r = \pm\sqrt{b_{XY}b_{YX}}$
4. Both the lines of regression pass through the point $(\overline{x}, \overline{y})$, the sample mean.

Line of regression – The line of best fit – Principle of Least squares :

Consider the line of regression of Y on X. Let this straight line be given by $y = a + bx$ —— (1)

Now we have to determine the constants a & b s.t (1) gives, for each value of x, the best value (estimate for the average value) of y. This is done using the Least square Principle.

Accordingly,

Let $S_i = y_i - y$

$\Rightarrow S_i = y_i - (a + bx_i)$

$\therefore S = \Sigma S_i^2 = \Sigma (y_i - (a + bx_i))^2$

$\Rightarrow \frac{\partial S}{\partial a} = 0 \Rightarrow -2\Sigma(y - (a+bx)) = 0$

& $\frac{\partial S}{\partial b} = 0 \Rightarrow -2 \Sigma x (y - (a+bx)) = 0$

From which we get the normal equations for a & b as,

$$\Sigma y = na + b\Sigma x \quad — (2)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad — (3)$$

(÷)ing (2) by $n \Rightarrow \bar{y} = a + b\bar{x}$

$\Rightarrow$ Thus $(\bar{x}, \bar{y})$ i.e. means of x & y lie on (1)

Shifting the origin to $(\bar{x}, \bar{y})$, (3) takes the form

$$\Sigma (x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2$$

But $a\Sigma(x - \bar{x}) = 0$ (why ?)

$\therefore \Rightarrow b = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\Sigma (x - \bar{x})^2}$

or $b = \frac{cov(x,y)}{\sigma_x^2}$

$b = r\frac{\sigma_y}{\sigma_x}$ $\left(\because r = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}\right)$

which is the slope of the line of regression of y on x thus the line of best fit becomes (since the line of regression passes through the point $(\bar{x}, \bar{y})$ with slope $b = r\frac{\sigma_y}{\sigma_x}$)

$\Rightarrow y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$

---

is the equation of the line of regression of y on x.

Its slope is called the coefficient of regression of y on x and is denoted by b_{yx}.

Interchanging x & y, i.e. starting with the equation $x = a + by$ and proceeding similarly we obtain the equation to the line of regression of x on y as

$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})$

Problems

1. Obtain the line of regression of Y on X for the following data and estimate the most probable value of Y when X is 70.

Item No	x	y
1	40	2.5
2	70	6.0
3	50	4.5
4	60	5.0
5	80	4.5
6	50	2.0
7	90	5.5
8	40	3.0
9	60	4.5
10	60	3.0

Solution:

The line of regression of Y on X is given by:

$(Y - \overline{Y}) = r_{xy}\frac{\sigma_Y}{\sigma_X}(X - \overline{X})$

Item No	X	Y	d_x=x-60	d_y=y-4.5	d_x^2	$d_x d_y$
1	40	2.5	-20	-2	400	40
2	70	6.0	10	1.5	100	15
3	50	4.5	-10	0	100	0
4	60	5.0	0	0.5	0	0
5	80	4.5	20	0	400	0
6	50	2.0	-10	-2.5	100	25
7	90	5.5	30	1.0	400	30
8	40	3.0	-20	-1.5	400	300
9	60	4.5	0	0	0	0
10	60	3.0	0	-1.5	0	0
			$\sum d_x = 0$	$\sum d_y = -4.5$	$\sum d_x^2 = 2400$	$\sum d_x d_y = 140$

With A=60 and B=4.5 we have,

$$\bar{x} = A + \frac{\sum dx}{n} = 60 + 0 = 60 \text{ and } \bar{y} = B + \frac{\sum dy}{n} = 4.5 + \left(\frac{-4.5}{10}\right) = 4.05$$

$$\sigma_x^2 = \frac{\sum d_x^2}{n} - \left(\frac{\sum d_x}{n}\right)^2 \text{ and } \sigma_y^2 = \frac{\sum d_y^2}{n} - \left(\frac{\sum d_y}{n}\right)^2$$

$$r = \frac{n\sum d_x.d_y - (\sum d_x)(\sum d_y)}{\sqrt{[n\sum d_x^2 - (\sum d_x)^2]\,[n\sum d_y^2 - (\sum d_y)^2]}}$$

$$r.\frac{\sigma_y}{\sigma_x} = \frac{\sum dxdy - \frac{\sum dx \sum dy}{n}}{\sum d^2x - \frac{(\sum dx)^2}{n}}$$

$$= \frac{140-0}{2400-0} = \frac{140}{2400}$$

$$= 0.06$$

Thus, the required line of regression of Y on X is given by:

$$(Y - \bar{Y}) = r_{xy}\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$Y - 4.05 = 0.06(X - 60)$$

$$\Rightarrow Y = 0.06X + 0.45$$

Which is the line of the equation of Y on X.

Now, when X = 70, we have

Y = 0.06*70+0.45 = 4.65

Thus Y=4.65 is the most probable estimated value of Y when X=70.

2. Obtain the line of regression of X and Y for the above data

Solve it!

3. The regression equations of two variables X and Y are X=0.7Y+5.2 and Y=0.3X+2.8. Find the means of the variables and the correlation coefficient.

Solution:

Since both the lines of regression passes through the point $(\overline{X}, \overline{Y})$ we have,

$\overline{X} = 0.7\overline{Y} + 5.2$ and $\overline{Y} = 0.3\overline{X} + 2.8$

Solving we get,

$\overline{X} = 9.06$

$\overline{Y} = 5.518$

Now, regression coefficient of Y on X is: b_{yx}=0.3

regression coefficient of X on Y is: b_{xy}=0.7

$\therefore r = \sqrt{b_{yx}b_{xy}} = \sqrt{0.3 * 0.7} = \sqrt{0.21} = 0.46$

4. The heights of 12 fathers and sons are given below:

Height of father x	165	160	170	163	173	158	178	168	173	170	175	180
Height of son y	173	168	173	165	175	168	173	165	180	170	173	178

Obtain the two regression lines and hence obtain r, the correlation coefficient

Solve it!

5. The equations of two lines of regression are 4X+3Y+7=0 and 3X+4Y+8=0. Find (i) the means of the variables X and Y

(ii) the Regression coefficients b_{yx} and b_{xy}

(iii) the Correlation coefficient r, between X and Y.

Solve it!

Multiple-Linear Regression:

In many of the real life situations it may happen that the dependent variable Y can more adequately be predicted when there are more than one independent variable, say, $X_1, X_2, X_3 \ldots X_k$. Typically one may have a relationship of the type:

$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$

One can obtain the coefficients a, b_1, b_2, …, b_k, as earlier, by using the Principle Least Squares and hence obtain the line of regression. This is called the multiple regression.

Non-linear Regression:

Quite often it is observed while using linear regression that the estimated (predicted) values of the dependent variable produce poor results. One reason behind this could be due to the fact that the variables may be far from being linearly related but a Curvi-linear relationship may be more appropriate.

For example, instead of $Y = a + bX$, a relationship of second degree, say of the type,

$Y = a + bX + cX^2$ may be more appropriate. By using Principle of Least Squares, one may obtain the coefficients a, b and c and hence arrived the regression equation which is non-linear.