Regression Lines – $(\bar{x}, \bar{y})$ and Slope

1. Equation of straight line ***y* on *x***:

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

where, $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

2. Equation of straight line ***x* on *y***:

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

where, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Note: $b_{xy} \neq b_{yx}$

Regression Lines – Least square method

Consider the equation of a straight line ***y* on *x***:

$$y = \boldsymbol{b}x + \boldsymbol{a}$$

Normal equations are

$$\sum y = \boldsymbol{b}\sum x + \boldsymbol{a}n$$

$$\sum yx = \boldsymbol{b}\sum x^2 + \boldsymbol{a}\sum x$$

Solving the above two equations we get the values of $\boldsymbol{a}$ and $\boldsymbol{b}$

$$\boxed{y = \boldsymbol{b}x + \boldsymbol{a}}$$

Regression Lines – Least square method

Consider the equation of a straight line ***x* on *y***:

$$x = \boldsymbol{d}y + \boldsymbol{c}$$

Normal equations are

$$\sum x = \boldsymbol{d}\sum y + \boldsymbol{c}n$$

$$\sum xy = \boldsymbol{d}\sum y^2 + \boldsymbol{c}\sum y$$

Solving the above two equations we get the values of ***c*** and ***d***

$$\boxed{x = \boldsymbol{d}y + \boldsymbol{c}}$$

Problems

1. From the following data obtain the two regression equations:

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

Solution:

Computation of Regression Equations

X	Y	XY	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\Sigma X = 30$	$\Sigma Y = 40$	$\Sigma XY = 214$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 340$

Regression equation of Y on X is given by

$$Y = a + bX \quad \text{... (i)}$$

Following two normal equations are required to be solved to determine the values of a and b.

$$\Sigma Y = Na + b\Sigma X$$

$$\Sigma XY = a\,\Sigma X + b\Sigma X^2$$

Substituting the values, we get $40 = 5a + 30b$ ... (ii)

$214 = 30a + 220b$ ... (iii)

Multiplying equation (ii) by 6, $240 = 30a + 180b$ ... (iv)

$214 = 30a + 220b$ ... (v)

Subtracting equation (v) from (iv), we get

$$-40b = 26 \quad \Rightarrow \quad b = -0.65$$

Substituting the value of b in equation (ii), we have

$$40 = 5a + 30\,(-0.65) \;\Rightarrow\; 5a = 40 + 19.5 = 59.5 \quad \Rightarrow \; a = 11.9$$

Putting the values of a and b in equation (i), we get the regression equation of Y on X as

$$Y = 11.9 - 0.65X$$

Regression equation of X on Y is given by

$$X = a + bY \qquad \text{... (vi)}$$

and the two normal equations are:

$$\Sigma X = Na + b\Sigma Y$$

$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get $\quad 30 = 5a + 40b \qquad$ ... (vii)

$$214 = 40a + 340b \qquad \text{... (viii)}$$

Multiplying equation (vii) by 8, $\quad 240 = 40a + 320b \qquad$ ... (ix)

$$214 = 40a + 340b \qquad \text{... (x)}$$

Subtracting equation (x) from (ix), we get

$$-20b = 26 \quad \Rightarrow \quad b = -1.3$$

Substituting the value of b in equation (vii), we have

$$30 = 5a + 40\,(-1.3) \quad \Rightarrow \quad 5a = 30 + 52 = 82 \quad \Rightarrow \quad a = 16.4$$

Putting the values of a and b in equation (vi), we get the regression equation of X on Y as

$$X = 16.4 - 1.3Y.$$

2. A study of prices at Chennai and Vellore gave the following data:

	Chennai	Vellore
Mean	19.5	17.75
S.D.	1.75	2.5

Also the **coefficient of correlation** between the two **is 0.8**. Estimate the most likely price of rice **(i)** at Chennai corresponding to the price of 18 at Vellore, and **(ii)** at Vellore corresponding to the price of 17 at Chennai.

Let the prices of rice at Chennai and Vellore be denoted by X and Y respec-tively. Then from the data,

$$\bar{x} = 19.5,\ \bar{y} = 17.75,\ \sigma_X = 1.75,\ \sigma_Y = 2.5 \text{ and } r_{XY} = 0.8.$$

Regression line of X on Y is

$$x - \bar{x} = \frac{r\,\sigma_X}{\sigma_Y}(y - \bar{y})$$

i.e.,
$$x - 19.5 = \frac{0.8 \times 1.75}{2.5}(y - 17.75)$$

$\therefore$ When $y = 18$,

$$x = 19.5 + \frac{0.8 \times 1.75}{2.5} \times 0.25$$

$$= 19.64$$

Regression line of Y on X is

$$y - \bar{y} = \frac{r\,\sigma_Y}{\sigma_X}(x - \bar{x})$$

i.e.,
$$y - 17.75 = \frac{0.8 \times 2.5}{1.75}(x - 19.5)$$

$\therefore$ When $x = 17$,

$$y = 17.75 + \frac{0.8 \times 2.5}{1.75} \times (-2.5)$$

$$= 14.89$$

3. For two variables X and Y, we have two regression lines 2X-7Y+3=0 and aX+bY+6=0 for some constants a and b. Suppose we know that mean of X is 2 and mean of Y is 1. Also the second regression line passes through (4, 5). Find the correlation coefficient of these two lines.

Given: Regression Lines

(x) $2x-7y+3=0$

$\Rightarrow 2x-7y=-3$ — (1)

(y) $ax+by+6=0$

$\Rightarrow ax+by=-6$ — (2)

Also, $(\bar{x}, \bar{y}) = (2, 1)$

It is given that eqn (2) passes through the point (4,5).

We know that, the regression lines passes through the point $(\bar{x}, \bar{y})$.

$\Rightarrow$ Eqn (2) passes through $(\bar{x}, \bar{y})$ also.

Thus, (2) $\Rightarrow$ $ax + by = -6$.

Point (2, 1) passing through eqn (2) $\Rightarrow$ $2a + b = -6$. ←(3)

Point (4, 5) passing through Eqn (2) $\Rightarrow$ $4a + 5b = -6$. ←(4)

Now, $\left.\begin{array}{l} 2a + b = -6 \\ 4a + 5b = -6. \end{array}\right\}$ $a = -4$

$b = 2$

Now, eqn. (2) $\Rightarrow -4x + 2y = -6$.

eqn. (1) $\Rightarrow 2x - 7y = -3$

We know that any one of the above equation have to be y on x and the other one have to be x on y.

We need to find out by "trial & error" method.

Let (1) be x on y

Then, $2x = 7y - 3$.

$$x = \underbrace{\left(\tfrac{7}{2}\right)}_{\to\, b_{xy}} y - 3/2$$

Let (2) be y on x.

Then, $+2y = \overset{4x}{\cancel{2xy}} - 6 \Rightarrow y = \overset{b_{yx}}{2x} - 3/2$

W.K.T, $r = \sqrt{b_{yx} \cdot b_{xy}}$ $[0 \le r^2 \le 1]$.

$$r^2 = b_{yx} \cdot b_{xy}$$

$$= (2)\left(7/2\right) \Rightarrow r^2 = 7 > 1$$

$\cancel{r = \sqrt{7}}$ Thus, our assumption is wrong.

Now, let ① be y on x.

$$-7y = -2x - 3$$

$$y = \left(\tfrac{2}{7}\right)x + 3$$

and

let ② be x on y.

$$-4x = -2y - 6$$

$$x = \left(\tfrac{1}{2}\right)y + \tfrac{3}{2}$$

Now,

$$r^2 = b_{yx} \cdot b_{xy}$$

$$= \left(\tfrac{2}{7}\right)\left(\tfrac{1}{2}\right) = \tfrac{1}{7}.$$

$$r^2 = \tfrac{1}{7} < 1 \text{ (valid)}$$

$$\Rightarrow r = \sqrt{\tfrac{1}{7}} \Rightarrow \boxed{r = 0.37796}$$