TOPIC 2.9- **Linear Regression**

Regression

Regression is a mathematical measure of the average relationship between two or more variables interms of the original limits of the da[ta]

The line of regression of y on x is giv[en] by

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

where r is the correlation coefficient.

The line of regression of x on y is gn. by

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

Angle between two lines of regression

The angle ω between the two lines of regression is given by

$$\tan \omega = \frac{1-r^2}{r}\left(\frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right)$$

Note

1. If $r = 0$, we get. $\tan\omega = \infty$

$$\boxed{\omega = \frac{\pi}{2}}$$

$\therefore$ when $r = 0$, the lines of regression are perpendicular to each other.

Regression coefficients

Regression coefficient of y on x

$$r\frac{\sigma_y}{\sigma_x} = b_{yx} \longrightarrow (1)$$

Regression coefficient of x on y

$$r\frac{\sigma_x}{\sigma_y} = b_{xy} \longrightarrow (2)$$

From (1) and (2), we get

Correlation coefficient $\boxed{r = \pm\sqrt{b_{xy}\, b_{yx}}}$

Note

The regression coefficients b_{yx} and b_{xy} can be obtained by the following formula

$$b_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

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

(1) From the following data, find

(i) the two regression equations

(ii) the coefficient of correlation between the marks in Economics and Statistics

(iii) The most likely marks in statistics when marks in Economics are 30.

Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Here $\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$

$\bar{y} = \frac{\sum y}{n} = \frac{380}{10} = 38$

Coefficient regression of y on x is

$$b_{yx} = \frac{\sum (x+\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = -\frac{93}{140}$$

$$= -0.6643$$

x	y	$x-\bar{x}$ $x-32$	$y-\bar{y}$ $y-38$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-12
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
320	380	0	0	140	398	-93

Coefficient regression of x on y is

$$b_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(y-\bar{y})^2} = \frac{-93}{398}$$

$$= -0.2337$$

(i) The equation of the line of regression of x on y

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

$$x-32 = -0.2337(y-38)$$

$$\boxed{x = -0.2337y + 40.8806}$$

The equation of the line of regression of Y on X

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

$$y - 38 = -0.6643(x - 32)$$

$$\boxed{y = -0.6643x + 59.2576}$$

(ii) Coefficient of correlation

$$r = \pm\sqrt{b_{xy}\, b_{yx}}$$

$$= \pm\sqrt{(-0.2337)(-0.6643)}$$

$$\boxed{r = -0.394}$$

(iii) To find the most likely marks in statistics (Y) when marks in Economics (x) are 30.

$$y = -0.6643x + 59.2576$$

Put $x = 30$, $y = 39.3286$

$$\boxed{y \simeq 39}$$

(2) The equation of two regression lines are $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$. Find the mean values of x & y and the correlation coefficient between x and y.

Since both the line of regression passing through $(\bar{x}, \bar{y})$, we get

$$8\bar{x} - 10\bar{y} = -66 \longrightarrow (1)$$
$$40\bar{x} - 18\bar{y} = 214 \longrightarrow (2)$$

$$(1) \times 5 \Rightarrow \quad 40\bar{x} - 50\bar{y} = -330$$
$$\underline{40\bar{x} \mp 18\bar{y} = {}_{-}214}$$
$$-32\bar{y} = -544$$
$$\boxed{\bar{y} = 17}$$

$$8\bar{x} - 170 = -66$$
$$8\bar{x} = 170 - 66$$
$$8\bar{x} = 104$$
$$\boxed{\bar{x} = 13}$$

$(1) \Rightarrow 8x - 10y = -66$

$10y = 8x + 66$

$y = \frac{4}{5}x + \frac{33}{5}$

$\Rightarrow \boxed{b_{yx} = \frac{4}{5}}$

$(2) \Rightarrow 40x - 18y = 214$

$40x = 18y + 214$

$x = \frac{9}{20}y + \frac{107}{20}$

$\Rightarrow \boxed{b_{xy} = \frac{9}{20}}$

$\therefore$ The correlation coefficient $r = \pm\sqrt{b_{xy}\, b_{yx}}$

$$r = \pm\sqrt{\left(\frac{9}{20}\right)\left(\frac{4}{5}\right)} = \sqrt{9/25} = \frac{3}{5}$$

$$\boxed{r = 0.6}$$