Regression

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

LINES OF REGRESSION:

* The line of Regression of x on y is

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

(or)

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

* The line of Regression of y on x is

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

(or)

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

where b_{xy}, b_{yx} are repression coefficients.

Correlation coefficient $r_{xy} = \pm \sqrt{b_{xy} \times b_{yx}}$

Properties:

1. When b_{xy} and b_{yx} are +ve r is +ve
2. When b_{xy} and b_{yx} both are negative r is -ve
3. When $r = \pm 1$ the repression lines either parallel or coincide.
4. When $r = 0$, the repression lines are perpendicular to each other.

4. The regression lines always passes through $(\bar{x}, \bar{y})$

5. If one of the regression coefficient is greater than 1 and another one must be less than one.

6. The Geometric mean of the regression coefficient is correlation coefficient.

$$r_{xy} = \pm\sqrt{b_{xy} \times b_{yx}}$$

7. The arithematic mean of the regression coefficient is greater than correlation coefficient

i.e., $\frac{b_{xy} + b_{yx}}{2} > r_{xy}$

8. The regression coefficients are independent of change of origin but not of scale.

i.e., $b_{xy} = b_{uv}$ where $u = x - a$, only.

9. The angle between two lines of regression is

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

Proof: WKT $\theta = \tan^{-1}\left(\frac{m_2 - m_1}{1 + m_1 m_2}\right)$

Target equation of the straight line is

$$y - y_1 = \frac{-1}{m}(x - x_1)$$

Let the regression of y on x is

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

$$-\frac{1}{m} = r\frac{\sigma_y}{\sigma_x}$$

$$\Rightarrow m_1 = \frac{-\sigma_x}{r\sigma_y}$$

Let the regression of x on y is

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

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

$$\Rightarrow -\frac{1}{m} = \frac{\sigma_y}{r\sigma_x}$$

$$\Rightarrow m_2 = -\frac{r\sigma_x}{\sigma_y}$$

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

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

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

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

* The two regression lines are $8x - 10y + 66 = 0$ and $40x - 18y - 214 = 0$. Find $(\bar{x}, \bar{y})$. Given $V(x) = 9$, find $V(y)$, find r_{xy}.

Soln:

Gn $\quad 8x - 10y = -66$

$\quad\quad 40x - 18y = 214$

WKT the regression lines always passes thro $(\bar{x}, \bar{y})$.

$$8\bar{x} - 10\bar{y} = -66$$

$$40\bar{x} - 18\bar{y} = 214$$

on Solving $\quad \bar{x} = 13, \ \bar{y} = 17$

Assume $8x - 10y = -66$ as regression of x on y

$$8x = 10y - 66$$

$$x = \frac{10}{8}y - \frac{66}{8}$$

$$\Rightarrow b_{xy} = \frac{10}{8}$$

Assume $40x - 18y = 214$ as regression of y on x

$$-18y = -40x + 214$$

$$y = \frac{40}{18}x - \frac{214}{18}$$

$$\Rightarrow b_{yx} = \frac{40}{18}$$

Let $\quad r = \pm\sqrt{b_{xy} \times b_{yx}}$

$$= \pm\sqrt{\frac{10}{8} \times \frac{40}{18}}$$

$r \geq 1$ the assumption is wrong.

Let $40x - 18y = 214$ as regression of x on y

$$\Rightarrow 40x = 18y + 214$$

$$x = \frac{18}{40} y + \frac{214}{40}$$

$$\Rightarrow b_{xy} = \frac{18}{40} = \frac{9}{20}$$

Let $8x - 10y = -66$ as regression of y on x

$$-10y = -8x - 66$$

$$y = \frac{-8}{-10} x - \frac{66}{-10}$$

$$\Rightarrow b_{yx} = \frac{8}{10}$$

$$r = \pm \sqrt{b_{xy} \times b_{yx}} = \pm \sqrt{\frac{9}{20} \times \frac{8}{10}}$$

$$r = \pm 0.6$$

$$\boxed{r = 0.6}$$

Gn $V(x) = 9$, find $V(y)$

$$\Rightarrow \sigma_x = 3$$

Let $b_{xy} = r \frac{\sigma_x}{\sigma_y}$

$$\frac{18}{40} = (0.6) \frac{3}{\sigma_y}$$

$$\sigma_y = \frac{(0.6) \times 3 \times 40}{18}$$

$$\sigma_y = 4$$

$$\Rightarrow V(y) = 16$$

2. Out of two lines of regression $x+2y=5$, $2x+3y=8$

(i) Find $(\bar{x}, \bar{y})$ (ii) Which one is the regression of x on y

(iii) If $V(x)=12$, find $V(y)$.

Soln:

Since the regression lines always passes through $(\bar{x}, \bar{y})$.

$$\bar{x}+2\bar{y}=5$$

$$2\bar{x}+3\bar{y}=8$$

on solving $\bar{x}=1$, $\bar{y}=2$

Let us assume $x+2y=5$ is the regression of y on x.

$$2y=-x+5$$

$$y=\frac{-1}{2}x+\frac{5}{2}$$

$$\Rightarrow byx=\frac{-1}{2}$$

Let $2x+3y=8$ as the regression of x on y.

$$2x=-3y+8$$

$$x=\frac{-3}{2}y+\frac{8}{2}$$

$$\Rightarrow bxy=\frac{-3}{2}$$

Let $r=\pm\sqrt{bxy \times byx}=\pm\sqrt{\left(\frac{-3}{2}\right)\left(\frac{-1}{2}\right)}$

$$r=\pm\sqrt{\frac{3}{4}}=\pm 0.86$$

$$\Rightarrow r=-0.86$$

(ii) $2x+3y$ is the regression of x on y.

(iii) Gn $V(x) = 12$

wkt $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$$\therefore \quad -\frac{3}{2} = -0.86 \times \frac{\sqrt{12}}{\sigma_y}$$

$$\sigma_y = \frac{0.86 \times 3.46}{1.5} = 1.98 \simeq 2$$

$$\sigma_y = 2 \implies V(y) = 4.$$

* Obtain equations of lines of regression from the following data. Obtain the value of y when $x = 71$. Also obtain the value of x when $y = 70$.

x:	65	66	67	67	68	69	70	72
y:	67	68	65	68	72	72	69	71

Soln:

$$\bar{x} = \frac{\Sigma x}{n} = 68 = E(x) \qquad E(xy) = 4695$$

$$\bar{y} = \frac{\Sigma y}{n} = 69 = E(y)$$

$$\sigma_x = \sqrt{E(x^2) - (E(x))^2} = \sqrt{4628.5 - (68)^2} = 2.12$$

$$\sigma_y = \sqrt{E(y^2) - (E(y))^2} = \sqrt{4766.5 - (69)^2} = 2.34$$

$$r_{xy} = \frac{E(xy) - E(x)E(y)}{\sigma_x \sigma_y} = 0.6014$$

Regression of x on y is

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

$$x - 68 = (0.604) \frac{2.12}{2.34} (y - 69)$$

$$x - 68 = 0.5472 (y - 69) = 0.5472y - 37.75$$

$$x - 0.5472y = 30.25 \rightarrow (1)$$

Regression of y on x is

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

$$y - 69 = (0.604) \frac{2.34}{2.12} (x - 68)$$

$$y - 69 = 0.666x - 45.33$$

$$y - 0.666x = 69 - 45.33$$

$$y - 0.666x = 23.67 \rightarrow (2)$$

To find x when $y = 70$

put $y = 70$ in (1)

$$x - 0.5472(70) = 30.25$$

$$x - 38.29 = 30.25$$

$$\boxed{x = 68.54}$$

To find y when $x = 71$

put $x = 71$ in (2)

$$y - 0.666(71) = 23.67$$

$$\Rightarrow y = 70.956 \approx 71$$

* The regression equation of x on y is $3y - 5x + 108 = 0$. If the mean value of y is 44 and the $V(x) = \left(\frac{9}{16}\right)^{th} V(y)$ find the mean value of x. Also find the correlation coefficient

Soln:

Regression of x on y is $3y - 5x + 108 = 0$.

Gn $\bar{y} = 44$. Since the regression lines passes thro $(\bar{x}, \bar{y})$

$$3\bar{y} - 5\bar{x} + 108 = 0$$

$$3(44) - 5\bar{x} + 108 = 0$$

$$182 - 5\bar{x} + 108 = 0$$

$$240 - 5\bar{x} = 0$$

$$-5\bar{x} = -240$$

$$\bar{x} = 48.$$

Since $3y - 5x + 108 = 0$ is the regression of x on y

$$-5x = -3y - 108$$

$$x = \frac{3}{5}y + \frac{108}{5}$$

$$b_{xy} = \frac{3}{5}$$

Wkt $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

$V(x) = \frac{9}{16} V(y)$

$\Rightarrow \sigma_x = \frac{3}{4}\sigma_y$

$$\frac{3}{5} = r\frac{\frac{3}{4}\sigma_y}{\sigma_y}$$

$$\therefore \frac{3}{5} = r\frac{3}{4}$$

$$\Rightarrow r = \frac{4}{5} = 0.8$$

* The tanget angle between the lines of regression of Y on X and X on Y is 0.6 and $\sigma_x = \frac{1}{2}\sigma_y$. find r_{xy}.

Soln:

$$\tan\theta = 0.6$$

$$\sigma_y = 2\sigma_x$$

WKT $$\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

$$0.6 = \left(\frac{1-r^2}{r}\right) \frac{\sigma_x \cdot 2\sigma_x}{\sigma_x^2 + 4\sigma_x^2}$$

$$0.6 = \frac{1-r^2}{r} \cdot \frac{2\sigma_x^2}{5\sigma_x^2}$$

$$0.6 = \left(\frac{1-r^2}{r}\right)\frac{2}{5}$$

$$(0.6)5r = 2(1-r^2)$$

$$3r = 2 - 2r^2$$

$$2r^2 + 3r - 2 = 0$$

$$r = 0.5, -2$$

$$\boxed{r = 0.5}$$

* If X and Y are jointly distributed with correlation $r_{xy} = \frac{1}{2}$, $\sigma_x = 2$, $\sigma_y = 3$. find the variance $2X - 4Y + 3$.

Soln:

$$Var(2X-4Y+3) = V(2X-4Y) + V(3)$$

$$= 4V(X) + 16V(Y) - 2(2)(4)\,Cov(X,Y)$$

WKT $r_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y} \Rightarrow$

$Cov(x,y) = r_{xy}\, \sigma_x \sigma_y$;

$= \frac{1}{7} \cdot 7 \cdot 3$

$\Rightarrow Cov(x,y) = 3.$

$\therefore V(2x - 4y + 3) = 4(4) + 16(9) - 16(3) = 112$

* From the following data, find the most likely price at Madras corresponding to the price 70 at Bombay and that at Bombay corresponding to the price 68 at Madras.

	Madras	Bombay
Average Price	65	67
SD of Price	0.5	3.5

Standard difference blw the price at Madras and Bombay is 3.1.

Soln: WKT $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$

$r = \frac{(0.5)^2 + (3.5)^2 - (3.1)^2}{2(0.5)(3.5)}$

$= \frac{0.25 + 12.25 - 9.61}{3.5}$

$\Rightarrow r = 0.825$

Regression of x on y is

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

$$x - 65 = (0.825)\left(\frac{0.5}{3.5}\right)(y - 67)$$

$$x - 65 = (0.117)(y - 67)$$

When Bombay = 70 Madras = ?

$$x - 65 = (0.117)(70 - 67)$$

$$x - 65 = 0.351$$

$$\Rightarrow x = 65.351$$

Regression of y on x is

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

$$y - 67 = (0.825)\left(\frac{3.5}{0.5}\right)(x - 65)$$

$$y - 67 = (5.775)(x - 65)$$

When $x = 68$ (Madras); Bombay = ?

$$y = 84.325$$