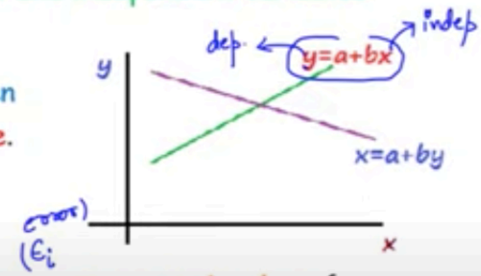
Unit 2 MTH302
Part II : Linear Regression with properties

Regression Lines:

A regression line is a graphic technique to show the functional relationship between two variables X and Y, i.e., dependent and independent variables.

It is a line which shows average relationship between two variables X and Y. Thus, this is a line of average.



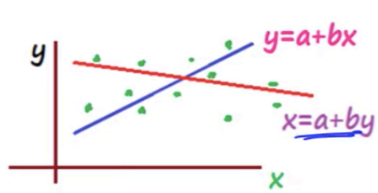
This is also called estimating lines, as it gives the average estimated value of dependent variable (Y) for any given value of independent variable (X).

Regression Equations/Estimating lines:

There are two algebraic expression of regression lines,

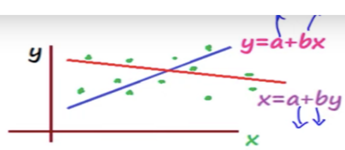
dependent indep

1) the regression equation of X on Y
$$X = a + bY$$
which shows the variation in the values of X for given changes in Y.



dep

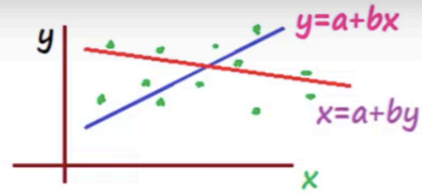
2) the regression equation of Y on X
$$Y = a + bX$$
which shows the variation in the values of Y for given changes in X.

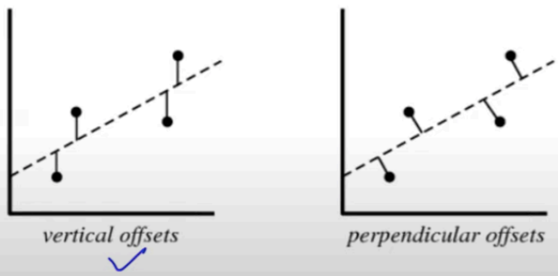


The objective of these regression lines is to fit the data on the lines.

For this, we need to estimate unknown parameters a and b

A mathematical procedure for finding the best-fitting curve to a given set of points by **minimizing the sum of the squares** of the offsets ("the residuals") of the points from the curve





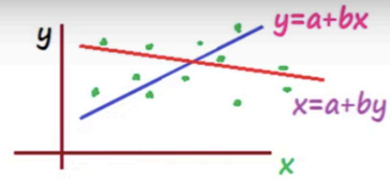
Note: In practice, the *vertical offsets* from a line are almost always minimized instead of the perpendicular offsets.

For Y on X; the regression equation is

$$Y = a + bX$$

For any given X, an estimated value Y_e of Y is

$$Y_e = a + bX$$



By principle of least squares, we minimize the residual errors i.e.,

$$E = \sum (Y - Y_e)^2$$

$$= \sum (Y - a - bX)^2$$

For minimization; we have

$$\frac{\partial E}{\partial a} = 0 \Rightarrow -2\sum(Y - a - bX) = 0$$

$$\frac{\partial E}{\partial b} = 0 \Rightarrow -2\sum(Y - a - bX)X = 0$$

$$\Rightarrow \sum(Y - a - bX) = 0 \ ;$$

$$\sum(XY - aX - bX^2) = 0$$

$$\boxed{\begin{aligned} &\Rightarrow \sum Y = na + b\sum X \ ; \\ &\sum XY = a\sum X + b\sum X^2 \end{aligned}}$$

These equations are called as NORMAL EQUATIONS.

Hence,

for Y on X; the regression equation is

$$Y = a + bX$$

Its normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Similarly,

a → 1
b → Y

For X on Y; the regression equation is

$$X = a + bY$$

$\sum XY = aY + bY^2$

The Normal equations are

$$\sum X = na + b\sum Y \checkmark$$

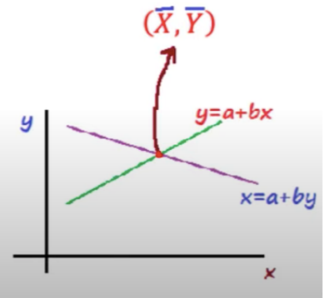
$$\sum XY = a\sum Y + b\sum Y^2 \checkmark$$

Note:

the point of intersection of the two-regression lines

$$Y = a + bX \text{ and } X = a + bY$$

gives the MEAN of the X and Y.



For Y on X line, regression line

$$Y = \underbrace{a + bX}, $$

slope is b, so call as regression coefficient of Y on X and is denoted as b_{yx}.

b_{yx}

Similarly,

For X on Y,

$$X = a + bY$$

b_{xy} represent the regression coefficient.

Expression of b_{xy} & b_{yx}

for Y on X; the regression equation is

$$Y = a + bX$$

Its normal equations are

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

After solving, we get "a" and "b"

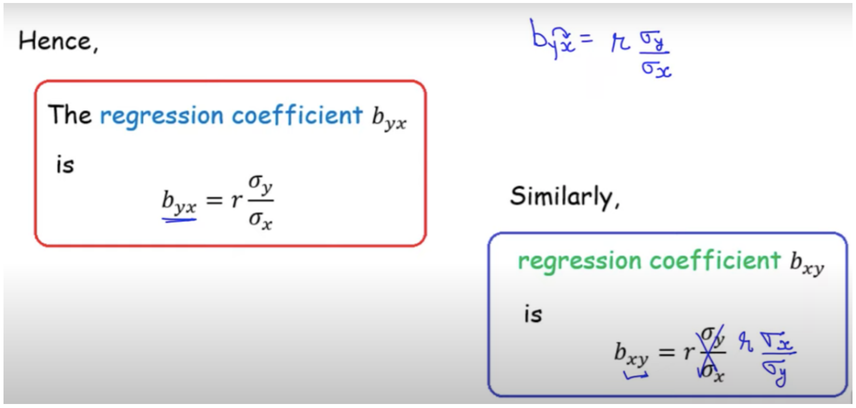
$$b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{\frac{\sum XY}{n} - \frac{\sum X}{n}\frac{\sum Y}{n}}{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

$$= \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2}$$

$$= \frac{cov(X,Y)}{Var(X)}$$

$$= r\frac{\sigma_y}{\sigma_x}$$



Properties of Regression Coefficients.

1.Correlation coefficient is the geometric mean between the regression coefficients.

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

The sign of r is the same as that of b_{xy} and b_{yx}

2. If one of the regression coefficients is greater than unity. the other must be less than unity.

3. Arithmetic. mean of the, regression coefficients is greater than the correlation coefficient r, provided r>0

$$\frac{1}{2}(b_{YX} + b_{XY}) \geq r$$

4. Regression coefficients are independent of the change of origin but not Of scale.

$$U = \frac{X - a}{h}, V = \frac{Y - b}{k}$$

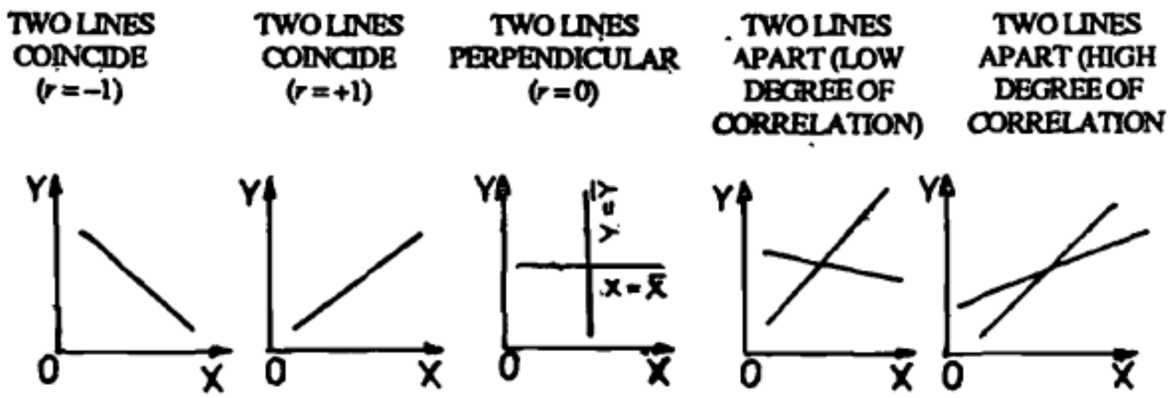
$$b_{YX} = \frac{k}{h} b_{VU}$$

And

$$b_{XY} = (h/k)\, b_{UV}$$

5.If the two variables are uncorrelated (r=0) , the lines of regression become perpendicular to each other in the case of perfect correlation $(r = \pm 1)$ positive or negative, the two lines of regression coincide.

But since both the lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel.



Case (i). ($r = 0$). If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Case (ii). ($r = \pm 1$). If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π.

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other

Example 1 *In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible: Variance of X = 1. The regression equations are 3x + 2y = 26 and 6x + y = 31. What were (i) the mean values of X and Y ? (ii) the standard deviation of Y ? and (iii) the correlation coefficient between X and Y ?*

Solution

(i) Since the lines of regression intersect at $(\bar{x}, \bar{y})$, we have $3\bar{x} + 2\bar{y} = 26$ and $6\bar{x} + \bar{y} = 31$

Solving these equations, we get $\bar{x} = 4$ and $\bar{y} = 7$.

(ii) Which of the two equations is the regression equation of Y on X and which one is the regression equation of X on Y are not known.

Let us tentatively assume that the first equation is the regression line of X on Y and the second equation is the regression line of Y on X. Based on this assumption, the first equation can be re-written as

$$x = -\frac{2}{3}y + \frac{26}{3} \qquad (1)$$

and the other as $\quad y = -6x + 31 \qquad (2)$

Then $b_{XY} = -\frac{2}{3}$ and $b_{YX} = -6$

$\therefore \quad r_{XY}^2 = b_{XY} \times b_{YX} = 4$

$\therefore \quad r_{XY} = -2$, which is absurd.

Hence our tentative assumption is wrong.

$\therefore$ The first equation is the regression line of Y on X and re-written as

$$y = -\frac{3}{2}x + 13 \qquad (3)$$

The second equation is the regression line of X on Y and re-written as

$$x = -\frac{1}{6}y + \frac{31}{6} \qquad (4)$$

Hence the correct $b_{YX} = -\frac{3}{2}$ and the correct $b_{XY} = -\frac{1}{6}$

$\therefore \quad r_{XY}^2 = b_{YX} \cdot b_{XY} = \frac{1}{4}$

$\therefore \quad r_{XY} = -\frac{1}{2} \qquad (\because$ both b_{YX} and b_{XY} are negative)

(iii) Now $\quad \dfrac{\sigma_Y^2}{\sigma_X^2} = \dfrac{b_{YX}}{b_{XY}} = \dfrac{-\frac{3}{2}}{-\frac{1}{6}} = 9$

$\therefore \quad \sigma_Y^2 = 9 \times \sigma_x{}^2 = 9$

$\therefore \quad \sigma_Y = 3$

2.

Example: The following results were declared in Physics and Mathematics in B.Tech examination. Find

(i) Regression lines

(ii) Regression coefficients

(iii) Estimate the value of Y when X = 40

(iv) Estimate the value of X when Y = 20.

	Scores in Physics (X)	Score in Mathematics (Y)
Mean	30	40
S.D.	10	20

Karl Pearson's coefficient of correlation between X and Y is 0.4.

Solution:

The regression line of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - 40 = b_{yx}(X - 30)$$

$$\Rightarrow Y - 40 = 0.8(X - 30)$$

$$\Rightarrow Y = 0.8X + 16$$

The regression line of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = 0.4\left(\frac{10}{20}\right) = 0.2$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = 0.4\left(\frac{20}{10}\right) = 0.8$$

The regression line of Y on X is

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$\Rightarrow Y - 40 = b_{yx}(X - 30)$$

$$\Rightarrow Y - 40 = 0.8(X - 30)$$

$$\Rightarrow Y = 0.8X + 16$$

When X = 40, then Y = 48

When Y = 20, then X = 26

The regression line of X on Y is

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$\Rightarrow X - 30 = 0.2(Y - 40)$$

$$\Rightarrow X = 0.2Y + 22$$

3.

Example: Estimate X when Y=10, if the two lines of regressions are

$$X = -\frac{1}{18}Y + \lambda; \; Y = -2x + \mu$$

$b_{yx} = -2$ $\quad b_{xy} = -\frac{1}{18}$

where (λ, μ) are unknown and the mean of the distribution is at $(-1, 2)$. Find r, λ, μ.

Solution: Since mean (-1, 2) passes through regression lines

$$\Rightarrow -1 = -\frac{2}{18} + \lambda \quad ; \quad 2 = 2 + \mu$$

$$\Rightarrow \lambda = -\frac{8}{9} \quad ; \; \mu = 0$$

$r = \sqrt{-\frac{1}{18} \times -2}$

$= -\frac{1}{3}$

Regression line of X on Y is

$$X = -\frac{1}{18}Y - \frac{8}{9}$$

When $Y = 10$;

then $X = -\frac{1}{18}(10) - \frac{8}{9}$

$= -\frac{13}{9}$

4.

Example: By using the following data, find the two lines of regression and Karl-Pearson's coefficient of correlation.

$$\sum X = 250\,; \sum Y = 300; \sum XY = 7900\,; \sum X^2 = 6500\,; \sum Y^2 = 10000, N = 10$$

Solution:

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{cov(X,Y)}{\sigma_X^2} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \frac{\frac{7900}{10} - \left(\frac{250}{10}\right)\left(\frac{300}{10}\right)}{\frac{6500}{10} - \left(\frac{250}{10}\right)^2} = 1.6$$

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 30 = 1.6(X - 25)$$

$$\Rightarrow Y = 1.6X - 10$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{cov(X,Y)}{\sigma_Y^2} = \frac{E(XY) - E(X)E(Y)}{E(Y^2) - (E(Y))^2} = \frac{\frac{7900}{10} - \left(\frac{250}{10}\right)\left(\frac{300}{10}\right)}{\frac{10000}{10} - \left(\frac{250}{10}\right)^2} = 0.4$$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 25 = 0.4(Y - 30)$$

$$\Rightarrow X = 0.4Y + 13$$

$r^2 = b_{xy} \,.\, b_{yx} = 0.4 \times 1.6 = 0.64$

$r = 0.8$

Standard Error of Estimate of *Y*

Although we use the regression line of Y on X to predict the value of Y corresponding to a specified value of X we may also use it to estimate the value of Y corresponding to an observed value of $X = x_i$, say. The value of Y estimated in this manner need not, in general, be equal to the corresponding observed value

of Y, namely, y_i. Hence the difference between Y and Y_E is called *the error of estimate of Y*. This error will vary from one observed value to the other and a random variable. The standard deviation of this RV $(Y - Y_E)$ is called *the standard error of estimate of Y* and denoted by S_Y.

The standard error of estimate of Y is S_Y

$$S^2_Y = = (1 - r^2_{\;}) \sigma^2_Y \text{ or } S_Y = \sqrt{1 - r^2_{\;}}\, \sigma_Y \tag{1}$$

Similarly, the standard error of estimate of X, denoted by S_X is given by

$$S^2_X = (1 - r^2_{\;}) \sigma^2_X \text{ or } S_X = \sqrt{1 - r^2_{\;}}\, \sigma_X \tag{2}$$

Example 5 *Find the standard error of estimate of Y on X and of X on Y from the following data:*

X:	1	2	3	4	5
Y:	2	5	9	13	14

Solution

x	y	x^2	y^2	xy
1	2	1	4	2
2	5	4	25	10
3	9	9	81	27
4	13	16	169	52
5	14	25	196	70
15	43	55	475	161

$$r_{XY} = \frac{n\Sigma xy - \Sigma x \cdot \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

$$= \frac{5 \times 161 - 15 \times 43}{\sqrt{\left\{5 \times 55 - (15)^2\right\}\left\{5 \times 475 - (43)^2\right\}}}$$

$$= \frac{160}{\sqrt{50 \times 526}} = 0.9866$$

$$\sigma^2{}_X = \frac{1}{n}\Sigma x^2 - \left(\frac{1}{n}\Sigma x\right)^2$$

$$= \frac{1}{5} \times 55 - \left(\frac{1}{5} \times 15\right)^2 = 2$$

$\therefore$ $$\sigma_X = 1.4142$$

$$\sigma^2{}_Y = \frac{1}{n}\Sigma y^2 - \left(\frac{1}{n}\Sigma y\right)^2$$

$$= \frac{1}{5} \times 475 - \left(\frac{1}{5} \times 43\right)^2$$

$$= 21.04$$

$\therefore$ $$\sigma_Y = 4.5869$$

$$S_Y = \sqrt{1 - r_{XY}^2} \cdot \sigma_Y = \sqrt{1 - (0.9866)^2} \times 4.5869$$

$$= 0.7484$$

$$S_X = \sqrt{1 - r_{XY}^2} \cdot \sigma_X = \sqrt{1 - (0.9866)^2} \times 1.4142$$

$$= 0.2307$$

Example: In partially destroyed laboratory record relating to correlation data, the following results are legible,

$\sigma_x^2 = 9$, Regression equations $8X - 10Y + 66 = 0;\ 40X - 18Y = 214$

Find (i) identify which line is of Y on X and X on Y (ii) mean of X and Y
(iii) S.D. of Y (iv) co-efficient of correlation between X and Y.

Solution: (i) Regression equations are

$$8X - 10Y + 66 = 0; \quad 40X - 18Y = 214$$

$$\Rightarrow Y = \frac{8}{10}X + \frac{66}{10} \quad ; \quad X = \frac{18}{40}Y + \frac{214}{40}$$

$$\Rightarrow b_{YX} = \frac{8}{10} \quad ; \quad b_{XY} = \frac{18}{40}$$

Thus,

Regression line of Y on X is $Y = \frac{8}{10}X + \frac{66}{10}$

Regression line of X on Y is $X = \frac{18}{40}Y + \frac{214}{40}$

(ii) Mean is the point of intersection of regression lines

$$8X - 10Y + 66 = 0;$$

$$40X - 18Y = 214$$

On solving, we get X = 13 & Y = 17

Thus, E(X) = 13 and E(Y) = 17

(iii) Correlation coefficient is

$$r = \sqrt{b_{XY} \times b_{YX}}$$

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

$$= 0.6$$

Find σ_Y.

Using $b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

$$\Rightarrow \frac{8}{10} = 0.6\frac{\sigma_Y}{3}$$

$$\Rightarrow \boxed{\sigma_Y = 4}$$

7

Example: If the two regression coefficients are 0.8 and 1.2, what would be the value of coefficient of correlation?

Solution: Given that $b_{xy} = {}^{+}0.8$; $b_{yx} = {}^{+}1.2$

b_{yx} ; b_{xy} & r have same sign

$$r = \sqrt{b_{xy} \times b_{yx}}$$

$$\pm \quad = \sqrt{0.8 \times 1.2}$$

$$= 0.98$$

8

Example: Find the coefficient of correlation from the following two regression equations: 3Y-2X-10=0 and 2Y-50-X=0. Also, estimate the value of Y when X=0.

Solution: The regression lines are

$$3Y - 2X - 10 = 0 \text{ and } 2Y - 50 - X = 0$$

Target

$$r = \sqrt{b_{XY} \times b_{YX}}$$

X on Y   Y on X

Y on X $\leftarrow$ $Y = \frac{2}{3}X + \frac{10}{3}$; $X = 2Y - 50 \rightarrow$ X on Y

$b_{YX} = \frac{2}{3}$; $b_{XY} = 2$

$\Downarrow$

$r = \sqrt{4/3} \notin [-1,1]$

$= 1.... >$

$\notin [-1,1]$

The regression lines are

$$3Y - 2X - 10 = 0 \text{ and } 2Y - 50 - X = 0$$

$$\Rightarrow X = \frac{3}{2}Y - 5 \quad ; \quad Y = \frac{1}{2}X + 25$$

$$\Rightarrow b_{XY} = \frac{3}{2} \quad ; \quad b_{YX} = \frac{1}{2}$$

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{\frac{3}{2} \times \frac{1}{2}}$$

$$= \frac{\sqrt{3}}{2} = 0.87$$

When X=0 then using regression line of Y on X, we get

$$Y = \frac{1}{2}(0) + 25 = 25$$

9

Example: The regression coefficient of regression equation of X on Y is 2.4 and of Y on X is 0.8. Are the regression coefficients consistent?

Solution: Given that $b_{XY} = 2.4$; $b_{YX} = 0.8$

Now, $r = \sqrt{b_{XY} \times b_{YX}}$

$= \sqrt{2.4 \times 0.8}$

$= \sqrt{1.92}$

$= 1.3856 \notin [-1, 1]$

Thus, the given regression coefficients are not consistent.

10

Example: Find the mean values of the variables X and Y and correlation coefficient for the following regression equations

$Y = \frac{1}{2}x + 25$ ← $2Y - X - 50 = 0$; $3Y - 2X - 10 = 0$ $X = \frac{3}{2}Y - \frac{10}{2}$

$b_{xy} = 3/2$

Solution: Mean is the point of intersection of regression lines

$b_{yx} = 1/2$

$2Y - X - 50 = 0$; $3Y - 2X - 10 = 0$

On solving, we get

$X = 130$ & $Y = 90$

Thus, $E(X) = 130$ and $E(Y) = 90$

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

$= \sqrt{\frac{1}{2} \times \frac{3}{2}}$

$= \sqrt{\frac{3}{4}} \in [-1, 1)$

$= \frac{\sqrt{3}}{2}$

11

Example: For 50 students of a class the regression equation of marks in Statistics (X) on marks in Mathematics (Y) is $3Y - 5X + 180 = 0$. The mean mark in mathematics is 44 and variance of marks in Statistics is 9/16th of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between the marks in two subjects.

Solution: Given that E(Y) = 44 ; $\sigma_X^2 = \frac{9}{16}\sigma_Y^2$

To find $E(X)$ and r

Since mean of X and Y passes through regression lines

Thus, $3\bar{Y} - 5\bar{X} + 180 = 0$

$\Rightarrow 3(44) - 5\bar{X} + 180 = 0$

$\Rightarrow \bar{X} = 62.4$

Regression line of X on Y is

3Y - 5X + 180 = 0

$\Rightarrow X = \frac{3}{5}Y + 36$

Thus, $b_{XY} = \frac{3}{5}$

$\Rightarrow r\frac{\sigma_X}{\sigma_Y} = \frac{3}{5}$

$\Rightarrow r\left(\frac{3}{4}\right) = \frac{3}{5}$

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

12

Example: Consider the following information about series X and Y. The coefficient of correlation between X and Y is +0.8. Find out the most probable value of Y if X is 70 and most probable value of X if Y is 90.

	Series X	Series Y
Mean	18	100
S.D.	14	20

Solution: When X = 70; then

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

Regression line of Y on X is

$Y - \bar{Y} = b_{YX}(X - \bar{X})$

$\Rightarrow Y - 100 = 0.8\left(\frac{20}{14}\right)(70 - 18)$

$\Rightarrow Y = 159.28$

When Y = 90, then

$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

Regression line of X on Y is

$X - \bar{X} = b_{XY}(Y - \bar{Y})$

$\Rightarrow X - 18 = 0.8\left(\frac{14}{20}\right)(90 - 100)$

$\Rightarrow X = 12.40$

13

Example: Given that the means of X and Y are 65 and 67, their standard deviations are 2.5 and 3.5 respectively and the correlation coefficient between them is 0.8.

(i) Write down the regression lines.

(ii) Obtain the best estimate of X when Y=70.

(iii) Using the estimated value of X as the given value of X, estimate the corresponding value of Y.

Solution: Given that $E(X) = 65$; $E(Y) = 67$; $\sigma_X = 2.5$; $\sigma_Y = 3.5$; $r = 0.8$

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

(i) Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 67 = 0.8\left(\frac{3.5}{2.5}\right)(X - 65)$$

$$\Rightarrow Y = 1.12X - 5.8$$

$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 65 = 0.8\left(\frac{2.5}{3.5}\right)(Y - 67)$$

$$\Rightarrow X = 0.571Y + 26.743$$

Regression line of Y on X is

$$\Rightarrow Y = 1.12X - 5.8$$

Regression line of X on Y is

$$\Rightarrow X = 0.571Y + 26.743$$

(ii) When Y = 70, then value of X is

$$X = 0.571(70) + 26.743$$
$$= 66.713$$

(iii) When X = 66.713, then Y is

$$Y = 1.12(66.713) - 5.8$$
$$= 68.92$$

14

Example: The correlation coefficient between X and Y variables is 0.60. If $\sigma_X = 1.5$, $\sigma_Y = 2.0, \bar{X} = 10,\ \bar{Y} = 20$, find the equations of the regression lines (i) Y on X (ii) X on Y.

Solution: Given that $r = 0.6$; $\sigma_X = 1.5$; $\sigma_Y = 2$; $\bar{X} = 10,\ \bar{Y} = 20$

$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 20 = 0.6\left(\frac{2}{1.5}\right)(X - 10)$$

$$\Rightarrow Y = \frac{4}{5}X + 12$$

$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 10 = 0.6\left(\frac{1.5}{2}\right)(Y - 20)$$

$$\Rightarrow X = 0.45Y + 1$$

15

Example: Find out σ_Y and r from the following data: $3X = Y$; $4Y = 3X$; $\sigma_X = 2$

Solution:

S.D

Y on X

X on Y

$Y = 3X$ $\quad X = \frac{4}{3}Y$

$b_{YX} = 3$ $\quad b_{XY} = \frac{4}{3}$

$r = \sqrt{3 \times \frac{4}{3}}$

$= 2 \notin [-1, 1]$

The regression lines are

$$3X = Y; \;\; 4Y = 3X$$

Regression line of Y on X

is $Y = \frac{3}{4}X$

$\Rightarrow b_{YX} = \frac{3}{4}$

Hence,

$$\boxed{\begin{aligned} r &= \sqrt{b_{XY} \times b_{YX}} \\ &= \sqrt{\frac{1}{3} \times \frac{3}{4}} \\ &= \frac{1}{2} \end{aligned}}$$

Regression line of X on Y

is $X = \frac{1}{3}Y$

$\Rightarrow b_{XY} = \frac{1}{3}$

For σ_Y

$$\boxed{\begin{aligned} b_{YX} &= r\frac{\sigma_Y}{\sigma_X} \\ \Rightarrow \frac{3}{4} &= \frac{1}{2}\frac{\sigma_Y}{2} \\ \Rightarrow \sigma_Y &= 3 \end{aligned}}$$

16

Example: The following data about the sales and advertisement expenditure of a firm is given in Table. Coefficient of correlation between them is **0.9**

	Sales (in crores)	Advertisement Expenditure (in crores)
Mean	40	6
S.D.	10	1.5

(i) Estimate the likely sales from a proposed advertisement expenditure of 10 crores.

(ii) What should be the advertisement expenditure if the firm proposes a sales target of 60 crores?

Solution:

(i) Given that Y = 10; X = ?

Regression line of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

$$\Rightarrow X - 40 = 0.9\left(\frac{10}{1.5}\right)(10 - 6)$$

$$\Rightarrow X = 64 \text{ crores}$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

(ii) Given that X = 60; Y = ?

Regression line of Y on X is

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

$$\Rightarrow Y - 6 = 0.9\left(\frac{1.5}{10}\right)(60 - 40)$$

$$\Rightarrow Y = 8.7 \text{crores}$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

17

Example: Given that the regression equation of Y on X and X on Y respectively $Y = X$ and $4X - Y = 3$. Find the correlation coefficient between X and Y.

Solution: Given that

Regression line of Y on X is

Y=X

$$\Rightarrow b_{YX} = 1$$

Regression line of X on Y is

$$X = \frac{1}{4}Y + \frac{3}{4}$$

$$\Rightarrow b_{XY} = \frac{1}{4}$$

Hence,

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{1 \times \frac{1}{4}}$$

$$= \frac{1}{2}$$

18

Example: From the following data, $X = 0.854\,Y$; $Y = 0.89X$; $\sigma_X = 3$

Calculate (i) coefficient of correlation (ii) S.D. of Y

Solution: From the given lines, we can get

$$b_{YX} = 0.89; \quad b_{XY} = 0.854$$

$$r = \sqrt{b_{XY} \times b_{YX}}$$

$$= \sqrt{0.89 \times 0.854}$$

$$= 0.87$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

$$\Rightarrow 0.854 = 0.87\frac{3}{\sigma_Y}$$

$$\Rightarrow \sigma_Y = 3.06$$

19.

Example 5 *Given that $x = 4y + 5$ and $y = kx + 4$ are the regression lines of X on Y and of Y on X respectively, show that $0 \le k \le \frac{1}{4}$. If $k = \frac{1}{16}$, find the means of X and Y and r_{XY}.*

Solution From the given equations, we note that

$$b_{YX} = k \text{ and } b_{XY} = 4$$

$$r_{XY}^2 = b_{XY} \cdot b_{YX} = 4k$$

Since $0 \le r_{XY}^2 \le 1$, we get $0 \le 4k \le 1$

$\therefore$ $$0 \le k \le \frac{1}{4}.$$

When $$k = \frac{1}{16}, r_{XY}^2 = \frac{1}{4}$$

$\therefore$ $$r_{XY} = \pm \frac{1}{2}$$

But both b_{YX} and b_{XY} are positive.

$\therefore$ $$r_{XY} = \frac{1}{2}$$

When $k = \frac{1}{16}$, the regression equations become

$$x = 4y + 5 \qquad (1)$$

and $$y = \frac{1}{16}x + 4 \qquad (2)$$

Solving equations (1) and (2), we get

$$x = 28 \text{ and } y = 5.75$$

$\therefore$ $$\bar{x} = 28 \text{ and } \bar{y} = 5.75$$

Angle Between Two Lines of Regression

$$Y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \text{ and } X - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

Slopes of these lines are $r \cdot \frac{\sigma_Y}{\sigma_X}$ and $\frac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the angle between the two lines of regression then

$$\tan\theta = \frac{r\frac{\sigma_Y}{\sigma_X} \sim \frac{\sigma_Y}{r\sigma_X}}{1 + r\frac{\sigma_Y}{\sigma_X} \cdot \frac{\sigma_Y}{r\sigma_X}} = \frac{r^2 \sim 1}{r} \cdot \left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \qquad (\because r^2 \le 1)$$

$\therefore$ $$\theta = \tan^{-1}\left\{\frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \qquad \ldots(10.19)$$

Case (i). ($r = 0$). If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2}$

Example 6 *Find the angle between the two lines of regression. Deduce the condition for the two lines to be (i) at right angles and (ii) coincident.*

Solution The equations of the regression lines

are $$y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(x - \bar{x}) \qquad (1)$$

and $$x - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(y - \bar{y}) \qquad (2)$$

Slope of line (1) $= r\dfrac{\sigma_Y}{\sigma_X} = m_1$, say.

Slope of line (2) $= \dfrac{\sigma_Y}{r\sigma_X}$, m_2, say.

If θ is the acute angle between the two lines, then $\tan\theta = \dfrac{|m_1 - m_2|}{1 + m_1 m_2}$

$$= \frac{\left| r\dfrac{\sigma_Y}{\sigma_X} - \dfrac{\sigma_Y}{r\sigma_X} \right|}{1 + \dfrac{\sigma_Y^2}{\sigma_X^2}}$$

$$= \frac{\left| r - \dfrac{1}{r} \right| \sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

$$= \frac{(1 - r^2)}{|r|} \frac{\sigma_X \sigma_Y}{\sigma_X^2 + \sigma_Y^2}$$

The two regression lines are at angles when $\theta = \dfrac{\pi}{2}$, i.e., $\tan\theta = \infty$

i.e., $r = 0$

$\therefore$ When the linear correlation between X and Y is zero, the two lines of regression will be at right angles.

The two regression lines are coincident, when $\theta = 0$, i.e., when $\tan\theta = 0$

i.e., when $r = \pm 1$.

$\therefore$ When the correlation between X and Y is perfect, the two regression lines will coincide.

Properties of Regression coefficients with proof

Properties

- Value of $r = \sqrt{b_{XY} \times b_{YX}} \in [-1,1]$
  - b_{XY}, b_{YX}, r all have same sign.
- If one of regression coefficient is greater than 1 then other MUST be less than one.

etc....

1

Property: The geometric mean between the two regression coefficients is equal to the correlation coefficient, i.e., $r = \sqrt{b_{XY} \times b_{YX}}$

Proof: The regression coefficients are

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad ; \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

b_{xy} b_{yx}

On multiplying, we get

$$b_{XY} \times b_{YX} = r^2$$

$$\Rightarrow \quad r = \sqrt{b_{XY} \times b_{YX}}$$

Hence result.

Property 2: If one regression coefficient is greater than unity, then others will be lesser than unity.

Proof: We know that $r = \sqrt{b_{XY} \times b_{YX}}$

r is correlation coefficient

$$\Rightarrow b_{XY} \times b_{YX} = r^2 \leq 1$$

$$\Rightarrow b_{XY} \leq \frac{1}{b_{YX}}$$

If $b_{YX} > 1$, then $b_{XY} < 1$ and vice versa too.

Hence, the result.

Property 3: Both of the regression coefficients must have the same sign.

i.e., If b_{YX} is positive, b_{XY} will also be positive and it is true for vice versa.

Proof: We know that

$$b_{XY} = r\frac{\sigma_X}{\sigma_Y} \quad ; \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

On Dividing, we get

$$\frac{b_{XY}}{b_{YX}} = \frac{\sigma_X^2}{\sigma_Y^2} > 0$$

Thus, both the regression coefficients have same sign.

Property 4: The regression coefficients b_{YX} , b_{XY} and correlation coefficient r have the same sign.

Proof: We know that $b_{XY} = r\frac{\sigma_X}{\sigma_Y} \quad ; \quad b_{YX} = r\frac{\sigma_Y}{\sigma_X}$

$$\Rightarrow \frac{b_{XY}}{r} = \frac{\sigma_X}{\sigma_Y} > 0 \quad ; \quad \frac{b_{YX}}{r} = \frac{\sigma_Y}{\sigma_X} > 0$$

Since $\sigma_X, \sigma_Y > 0$,

Thus, nature of b_{XY}, b_{YX} and r have always same sign.

Hence, the result.

Example: The regression coefficient of regression equation of X on Y is $-2/3$ and of Y on X is $-3/4$. Find the correlation coefficient.

Solution: $b_{XY} = -\frac{2}{3} \quad ; \quad b_{YX} = -\frac{3}{4}$

We know that $r = \sqrt{b_{XY} \times b_{YX}}$

$$= \pm\sqrt{-\frac{2}{3} \times -\frac{3}{4}}$$

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

Property 5: The arithmetic means of both regression coefficients is equal to or greater than the coefficient of correlation, i.e., $\frac{b_{XY}+b_{YX}}{2} \geq r$

Proof: The regression coefficients are

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad ; \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y}$$

Now,

$$\frac{b_{XY} + b_{YX}}{2} = \frac{r}{2}\left(\frac{\sigma_X}{\sigma_Y} + \frac{\sigma_Y}{\sigma_X}\right)$$

$$= r\left(\frac{\sigma_X^2 + \sigma_Y^2}{2\sigma_X\sigma_Y}\right)$$

Since

$$\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y = (\sigma_X - \sigma_Y)^2 \geq 0$$

$$\Rightarrow \sigma_X^2 + \sigma_Y^2 \geq 2\sigma_X\sigma_Y$$

$$\Rightarrow \frac{\sigma_X^2 + \sigma_Y^2}{2\sigma_X\sigma_Y} \geq 1$$

$$\Rightarrow r\left(\frac{\sigma_X^2 + \sigma_Y^2}{2\sigma_X\sigma_Y}\right) \geq r$$

$$\Rightarrow \frac{b_{XY} + b_{YX}}{2} \geq r$$

Property 6: The regression coefficients are independent of the change of the origin, i.e., if $U = X \pm a$; $V = Y \pm b$ then $b_{UV} = b_{XY}$; $b_{VU} = b_{YX}$

Proof: By definition, $b_{UV} = r\frac{\sigma_U}{\sigma_V}$ where $r = r(U,V)$

We know that $r(X \pm a, Y \pm b) = r(X,Y)$, i.e., $r(U,V) = r(X,Y) = r$

$$\sigma_U^2 = Var(U) = Var(X \pm a) = Var(X) = \sigma_X^2$$

$$\sigma_V^2 = Var(V) = Var(Y \pm b) = Var(Y) = \sigma_Y^2$$

Hence, $b_{UV} = r\frac{\sigma_X}{\sigma_Y} = b_{XY}$

Similarly, $b_{VU} = b_{YX}$

Example: The regression coefficient of regression equation of X on Y is 0.4 and of Y on X is 1.6. Find the regression coefficients of $X + 3$ on $Y - 2$ and Y-2 on X+3.

Solution: Since $b_{XY} = 0.4$; $b_{YX} = 1.6$

Take $U = X + 3$; $V = Y - 2$, we get

$$b_{UV} = b_{XY} = 0.4$$

$$b_{VU} = b_{YX} = 1.6$$

Property 7: The regression coefficients are $\boxed{\text{not}}$ independent of the change of the scale. If $U = aX$; $V = bY$ then $b_{UV} \neq b_{XY}$; $b_{VU} \neq b_{YX}$

If x and y are multiplied by any constant, then the regression coefficient will change.

Proof: $b_{UV} = r\frac{\sigma_U}{\sigma_V}$ where $r = r(U, V)$

We know that $r(aX, bY) = \begin{cases} r(X,Y) & \text{if } ab > 0 \\ -r(X,Y) & \text{if } ab < 0 \end{cases}$

$$\begin{aligned} \sigma_U^2 &= Var(U) \\ &= Var(aX) \\ &= a^2 Var(X) \\ &= a^2 \sigma_X^2 \end{aligned}$$

$$\begin{aligned} \sigma_V^2 &= Var(V) \\ &= Var(bY) \\ &= b^2 Var(Y) \\ &= b^2 \sigma_Y^2 \end{aligned}$$

Hence, $b_{UV} = \begin{cases} r\frac{a\sigma_X}{b\sigma_Y} & ; \text{ if } ab > 0 \\ -r\frac{a\sigma_X}{b\sigma_Y} & ; \text{ if } ab < 0 \end{cases}$

$= \begin{cases} \frac{a}{b} b_{XY} & ; \text{ if } ab > 0 \\ -\frac{a}{b} b_{XY} & ; \text{ if } ab < 0 \end{cases}$

Similarly, $b_{VU} = \begin{cases} \frac{b}{a} b_{YX} & ; \text{ if } ab > 0 \\ -\frac{b}{a} b_{YX} & ; \text{ if } ab < 0 \end{cases}$

Example: The regression coefficient of regression equation of X on Y is 0.4 and of Y on X is 1.6. Find the regression coefficients of $3X$ on $2Y$ and 2Y on 3X.

Solution: Since $b_{XY} = 0.4$; $b_{YX} = 1.6$

Take $U = 3X$; $V = 2Y$, we get

b/a

$$
\begin{aligned}
b_{UV} &= \frac{3}{2} b_{XY} \\
&= \frac{3}{2}(0.4) \\
&= 0.6
\end{aligned}
$$

$$
\begin{aligned}
b_{VU} &= \frac{2}{3} b_{YX} \\
&= \frac{2}{3}(1.6) \\
&= 1.066
\end{aligned}
$$