2. 3. Covariance

If X and Y are random variables, then covariance between X and Y is defined as

$$Cov\ (X, Y) = E\{[X - E(X)][Y - E(Y)]\}$$

$$= E\{XY - XE(Y) - E(X)Y + E(X)E(Y)\}$$

$$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)$$

$$Cov\ (X, Y) = E(XY) - E(X).E(Y) \qquad ...(A)$$

If X and Y are independent then $E(XY) = E(X)E(Y)$ ...(B)

Sub (B) in (A), we get $Cov\ (X, Y) = 0$

Therefore, if X and Y are independent then $Cov\ (X, Y) = 0$

Correlation:

If the change in one variable affects a change in the other variable, the variables are said to be correlated.

Two types of correlations are Positive correlation, Negative correlation

Positive Correlation:

If the two variables deviate in the same direction

Eg: Height and Weight of a group of persons, Income and Expenditure.

Negative Correlation:

If the two variables constantly deviate in opposite directions.

Eg: Price and Demand of a commodity, the correlation between volume and pressure of a perfect gas.

Measurement of Correlation:

We can measure the correlation between the two variables by using Karl – Pearson's coefficient of correlation.

Karl - Pearson' s coefficient of correlation

Correlation coefficient between two random variables X and Y, usually denoted by

$$r(X,Y) = \frac{COV(X,Y)}{\sigma_X.\sigma_Y}$$

Where $COV(X,Y) = \frac{1}{n}\sum XY - \bar{X}\,\bar{Y}$

$$\sigma_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X^2}}\ , \bar{X} = \frac{\sum X}{n}$$

$$\sigma_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y^2}}\ , \bar{Y} = \frac{\sum Y}{n}$$

(n is the number of items in the given data)

Note:

1. Correlation coefficient may also be denoted by $\rho(X,Y) or\ \rho_{XY}$

2. If $\rho(X,Y) = 0$, We say that X and Y are uncorrelated.

3. Correlation coefficient does not exceed unity.

Note:

Types of correlation based on 'r'

Value of 'r'	Correlation is said to be
$r = 1$	Perfect and positive
$0 < r < 1$	Positive
$-1 < r < 0$	Negative
$r = 0$	Uncorrelated

Problems under Correlation

1. Calculate the correlation coefficient for the following heights (in inches) of father X and their sons Y.

X	**65**	**66**	**67**	**67**	**68**	**69**	**70**	**72**
Y	**67**	**68**	**65**	**68**	**72**	**72**	**69**	**71**

Solution:

X	Y	XY	X^2	Y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
$\sum(X) = 544$	$\sum(Y) = 552$	$\sum(XY) = 37560$	$\sum(X^2) = 37028$	$\sum(Y^2) = 38132$

$$\bar{X} = \frac{544}{8} = 68\ ,\ \bar{Y} = \frac{552}{8} = 69$$

$$\bar{X}\,\bar{Y} = 68\ X\ 69 = 4692$$

$$\sigma_X = \sqrt{\frac{1}{n}\sum X^2 - \overline{X^2}} = 2.121$$

$$\sigma_Y = \sqrt{\frac{1}{n}\sum Y^2 - \overline{Y^2}} = 2.345$$

$$r(X,Y) = \frac{COV(X,Y)}{\sigma_X.\sigma_Y} = \frac{\frac{1}{n}\sum XY - \bar{X}\bar{Y}}{\sigma_X.\sigma_Y} = \frac{\frac{1}{8}37560 - 4692}{2.121\ X\ 2.345} = 0.6031$$

2. Find the correlation coefficient between industrial production and export using the following data:

Production (X)	55	56	58	59	60	60	62
Export (Y)	**35**	**38**	**37**	**39**	**44**	**43**	**44**

Solution:

U	V	$U = X - 58$	$V = Y - 40$	UV	U^2	V^2
55	35	-3	-5	15	9	25
56	38	-2	-2	4	4	4
58	37	0	-3	0	0	9
59	39	1	-1	-1	1	1
60	44	2	4	8	4	16
60	43	2	3	6	4	9
62	44	4	4	16	16	16
		$\sum(U) = 4$	$\sum(V) = 0$	$\sum(UV) = 48$	$\sum(U^2) = 38$	$\sum(V^2) = 80$

Now $\quad \bar{U} = \frac{\sum U}{n} = \frac{4}{7} = 0.5714$

$$\bar{V} = \frac{\sum V}{n} = 0$$

$$Cov\ (U, V) = \frac{\sum UV}{n} - \bar{U}\bar{V} = 6.857$$

$$\sigma_X = \sqrt{\frac{1}{n}\sum U^2 - \overline{U^2}} = 2.2588$$

$$\sigma_Y = \sqrt{\frac{1}{n}\sum V^2 - \overline{V^2}} = 3.38$$

$$r\ (U, V) = \frac{COV\,(U,V)}{\sigma_U.\sigma_V} = \frac{\frac{1}{n}\sum UV - \bar{U}\bar{V}}{\sigma_U.\sigma_V} = 0.898$$

3. Two R.V.'S X and Y have joint p.d.f of $f(x, y) = \begin{cases} \frac{xy}{96} & 0 < x < 4,\ 1 < y < 5 \\ 0 & elsewhere \end{cases}$

Find (i) $E(X)$ (ii) $E(Y)$ (iii) $E(XY)$ (iv) $E(2X + 3Y)$ (v) $Var(X)$ (vi) $Var(Y)$ (vii) $Cov(X, Y)$

Solution:

(i) $E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x, y)\,dxdy$

$$= \int_1^5 \int_0^4 x\ \frac{xy}{96}\,dxdy$$

$$=\frac{1}{96}\int_1^5\int_0^4 x^2 y\,dx\,dy$$

$$= \frac{1}{96}\int_1^5 y\left(\frac{x^3}{3}\right)_0^4 dy$$

$$= \frac{64}{288}\int_1^5 y\,dy = \frac{2}{9}\left[\frac{y^2}{2}\right]_1^5 = \frac{24}{9}$$

(ii) $E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\,dx\,dy$

$$= \int_1^5\int_0^4 y\,\frac{xy}{96}\,dx\,dy$$

$$=\frac{1}{96}\int_1^5\int_0^4 xy^2\,dx\,dy$$

$$= \frac{1}{96}\int_1^5 y^2\left(\frac{x^2}{2}\right)_0^4 dy$$

$$= \frac{1}{96\,(2)}\int_1^5 y^2\,(4^2-0)\,dy$$

$$= \frac{16}{192}\left[\frac{y^3}{3}\right]_1^5 = \frac{124}{36}$$

$$\Rightarrow E(Y) = \frac{31}{9}$$

(iii) $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dx\,dy$

$$= \int_1^5\int_0^4 xy\left(\frac{xy}{96}\right)dx\,dy$$

$$= \frac{1}{96}\int_1^5\int_0^4 x^2y^2\,dx\,dy$$

$$= \frac{1}{96}\int_1^5 y^2 (\frac{x^3}{3})_0^4\, dy$$

$$= \frac{1}{96\,(3)} \int_1^5 y^2\, (4^3 - 0) dy$$

$$= \frac{64}{288} \int_1^5 y^2\, dy = \frac{2}{9}\left[\frac{y^3}{3}\right]_1^5 = \frac{248}{27}$$

$$\Rightarrow E(XY) = \frac{248}{27}$$

(iv) $E[2X + 3Y] = 2\,E[X] + 3\,E[Y]$

$$= 2\left(\frac{8}{3}\right) + 3(\frac{31}{9})$$

$$= \frac{16+31}{3} = \frac{47}{3}$$

(v) $Var(X) = E(X^2) - [E(X)]^2$

Now, $E(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2\, f(x,y) dxdy$

$$= \int_1^5 \int_0^4 x^2\, (\frac{xy}{96}) dxdy$$

$$= \frac{1}{96}\int_1^5 \int_0^4 x^3 y dxdy$$

$$= \frac{1}{96}\int_1^5 y(\frac{x^4}{4})_0^4\, dy$$

$$= \frac{1}{96\,(4)} \int_1^5 y\, (4^4 - 0) dy$$

$$= \frac{256}{384}\int_1^5 y\,dy = \frac{2}{3}\left[\frac{y2}{2}\right]_1^5 = \frac{24}{3} = 8$$

$$\Rightarrow E(X^2) = 8$$

$$Var\ (X\ ) = \mathrm{E}\ (X^2) - [E(X)]^2 = 8 - \left(\frac{8}{3}\right)^2 = \frac{8}{9}$$

$$\sigma_X^2 = \frac{8}{9} \Rightarrow \sigma_X = \frac{\sqrt{8}}{3}$$

$$E(Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2\ f(x,y)dxdy$$

$$= \int_1^5\int_0^4 y^2\ (\frac{xy}{96})dxdy$$

$$= \frac{1}{96}\int_1^5\int_0^4 xy^3dxdy$$

$$= \frac{1}{96}\int_1^5 y^3(\frac{x^2}{2})_0^4\,dy$$

$$= \frac{1}{96\ (2)}\int_1^5 y^3\ (4^2 - 0)dy$$

$$= \frac{16}{192}\int_1^5 y^3\ dy$$

$$= \frac{1}{12}\left[\frac{y^4}{4}\right]_1^5 = \frac{624}{48} = 13$$

$$\Rightarrow E(Y^2) = 13$$

(vi) $Var\ (Y\ ) = \mathrm{E}\ (Y^2) - [E(Y)]^2$

$= 13 - (\frac{31}{9})^2$

$= \frac{92}{81}$

$\sigma_Y^2 = \frac{92}{81} \Rightarrow \sigma_Y = \frac{\sqrt{92}}{9}$

(vii) $Cov\ (X, Y) = E\ (XY) - E(X) . E(Y)$

$= \frac{248}{27} - \frac{248}{27} = 0$

4. If the independent random variables X and Y have the variances 36 and 16 respectively, find the correlation coefficient between $X + Y$ and $X - Y$

Solution:

Given that $Var\ (X) = 36,\ Var\ (Y) = 16$. Since X and Y are independent,

$E(XY) = E(X) . E(Y)$

Let $U = X + Y$ and $V = X - Y$

$Var\ (U) = Var\ (X + Y)$

$= 1^2 Var\ (X) + 1^2 Var(Y)$

$(\because Var\ (a X + b Y) = a^2\ Var(X) + b^2\ Var\ (Y))$

$= 36 + 16 = 52$

$\sigma_U = \sqrt{52}$

$Var\ (V) = Var\ (X - Y)$

$= 1^2 Var\ (X) + (-1)^2 Var(Y)$

$(\because\ Var\ (a X + b Y) = a^2\ Var(X) + b^2\ Var\ (Y))$

$= 36 + 16 = 52$

$\sigma_V = \sqrt{52}$

$Cov\ (U,V) = E(UV) - E(U).E(V)$ ... (1)

$E(UV) = E\ [(X + Y)\ (X - Y)]$

$= E\ [X^2 - Y^2]$

$= E\ (X^2) - E(Y^2)$ . . . (2)

$E(U) = E(X + Y) = E(X) + E(Y)$ . . . (3)

$E(V) = E(X - Y) = E(X) - E(Y)$ . . . (4)

Substituting (2), (3), (4) in (1) we get

$Cov(U,V) = E(X^2) - E(Y^2) - [E(X) + E(Y)][E(X) - E(Y)]$

$= E\ (X^2) - E(Y^2) - [E(X)]^2 + [E(Y)]^2 - E(X)E(Y) + E(X)E(Y)$

$= \{E(X^2) - [E(X)]^2\} - \{E(Y^2) - [E(Y)]^2\}$

$= Var(X) - Var(Y)$

$$Cov\,(U,V) = 36 - 16 = 20$$

Hence $\rho(U,V) = \frac{Cov(U,V)}{\sigma_U \sigma_V} = \frac{20}{\sqrt{52\sqrt{52}}} = \frac{20}{52} = \frac{5}{13}$

5. If the joint pdf of (X,Y) is given by $f(x,y) = x + y\,, 0 \leq x, y \leq 1, find$

ρ_{XY}

Solution:

Given $f(x,y) = x + y, 0 \leq x, y \leq 1$

Now, $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x,y) dxdy$

$$= \int_1^1 \int_0^1 xy\,(x+y) dxdy$$

$$= \int_1^1 \int_0^1 (x^2y + xy^2) dxdy$$

$$= \int_1^1 \left[\frac{x^3y}{3} + \frac{x^2y^2}{2}\right]_0^1 dy$$

$$= \int_1^1 \left[\frac{y}{3} + \frac{y^2}{2}\right] dy$$

$$= \left[\frac{y^2}{6} + \frac{y^3}{6}\right]_0^1$$

$$= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

Marginal pdf of X is $f(x) = \int_0^1 f(x,y) dy$

$$= \int_0^1 (x+y)dy$$

$$= \left[xy + \frac{y^2}{2}\right]_0^1$$

$$= \left(x + \frac{1}{2}\right) - (0+0)$$

$$= \left(x + \frac{1}{2}\right)$$

Marginal pdf of Y is $f(y) = \int_0^1 f(x,y)dx$

$$= \int_0^1 (x+y)dx$$

$$= \left[\frac{x^2}{2} + xy\right]_0^1$$

$$= \left(\frac{1}{2} + y\right) - (0+0)$$

$$= \left(y + \frac{1}{2}\right)$$

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$= \int_0^1 x\left(x + \frac{1}{2}\right)dx$$

$$= \int_0^1 \left(x^2 + \frac{x}{2}\right)dx$$

$$= \left[\frac{x^3}{3} + \frac{x^2}{4}\right]_0^1$$

$$= \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

$$E(Y) = \int_{-\infty}^{\infty} y f(y) dy$$

$$= \int_0^1 y\left(y + \frac{1}{2}\right) dy$$

$$= \int_0^1 \left(y^2 + \frac{y}{2}\right) dy$$

$$= \left[\frac{y^3}{3} + \frac{y^2}{4}\right]_0^1$$

$$= \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx$$

$$= \int_0^1 x^2\left(x + \frac{1}{2}\right) dx$$

$$= \int_0^1 \left(x^3 + \frac{x^2}{2}\right) dx$$

$$= \left[\frac{x^4}{4} + \frac{x^3}{3}\right]_0^1$$

$$= \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

$$E(Y^2) = \int_{-\infty}^{\infty} y^2 f(y) dy$$

$$= \int_0^1 y^2 \left(y + \frac{1}{2}\right) dy$$

$$= \int_0^1 \left(y^3 + \frac{y^2}{2}\right) dx$$

$$= \left[\frac{y^4}{4} + \frac{y^3}{3}\right]_0^1$$

$$= \frac{1}{4} + \frac{1}{6} = \frac{5}{12}$$

$$Var(X) = E(X^2) - [E(X)]^2$$

$$= \frac{5}{12} - \left[\frac{7}{12}\right]^2 = \frac{11}{144}$$

$$\sigma_X{}^2 = \frac{11}{144}$$

$$\sigma_X = \frac{\sqrt{11}}{12}$$

$$Var(Y) = E(Y^2) - [E(Y)]^2$$

$$= \frac{5}{12} - \left[\frac{7}{12}\right]^2 = \frac{11}{144}$$

$$\sigma_Y{}^2 = \frac{11}{144}$$

$$\sigma_Y = \frac{\sqrt{11}}{12}$$

Correlation coefficient $\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \, \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sigma_X \, \sigma_Y}$

$$= \frac{\frac{1}{3} - \frac{7}{12} \times \frac{7}{12}}{\frac{\sqrt{11}}{12} \times \frac{\sqrt{11}}{12}} = -\frac{1}{11}$$

Rank Correlation:

$$r = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}$$

Where $d_i = x_i - y_i$

This formula is called Spearman's formula for Rank correlation.

Problems on Rank correlation:

1. Find the rank correlation from the following data.

X	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Rank in y	**4**	**3**	**1**	**2**	**6**	**5**	**7**

Solution:

X	Y	$d_i = x_i - y_i$	${d_i}^2$
1	4	-3	9
2	3	-1	1
3	1	2	4
4	2	2	4
5	6	-1	1

6	5	1	1
7	7	0	0
		$\sum d_i = 0$	$\sum {d_i}^2 = 20$

Rank correlation coefficient

$$r = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

$$r = 1 - \frac{6\times 20}{7(49-1)}$$

$$= 1 - \frac{120}{336} = 0.6429$$

2. The ranks of some 16 students in Mathematics and Physics are as follows. Calculate rank correlation coefficients for proficiency in Mathematics and Physics

Rank in Maths	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Rankin Physics	**1**	**10**	**3**	**4**	**5**	**7**	**2**	**6**	**8**	**11**	**15**	**9**	**14**	**12**	**16**	**13**

Solution:

X	Y	$d_i = x_i - y_i$	${d_i}^2$
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25
8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	9
		$\sum d_i = 0$	$\sum {d_i}^2 = 136$

Rank correlation coefficient

$$r = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)}$$

$$r = 1 - \frac{6\times 136}{16(256-1)}$$

$$= 1 - \frac{816}{4080} = 0.8$$

Regression:

Regression is a mathematical measure of the average relationship between two or more

variables in terms of the original limits of the data.

Lines of Regression

If the variables in a bivariate distribution are related we will find that the points in the scattered diagram will cluster around some curve called the curve of regression. If the curve is a straight line, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

The line of regression of Y on X is given by

$$\boldsymbol{y - \overline{y} = r.\frac{\sigma_Y}{\sigma_X}(x - \overline{x})}$$

Where r is the correlation coefficient, σ_Y and σ_X are standard deviations.

The line of regression of Y on X is given by

$$\boldsymbol{x - \overline{x} = r.\frac{\sigma_Y}{\sigma_X}(y - \overline{y})}$$

Note:

Both the lines of regression passes through the mean $(\bar{x}, \bar{y})$

Angle between two lines of regression

If the equations of lines of regression of Y on X and X on Y are

$$\boldsymbol{y - \overline{y} = r.\frac{\sigma_Y}{\sigma_X}(x - \overline{x})}$$

$$\boldsymbol{x - \overline{x} = r.\frac{\sigma_Y}{\sigma_X}(y - \overline{y})}$$

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

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

Problems on Regression

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**

Solution:

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
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	-24
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

Now $\bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32$

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

Coefficient of regression of Y on X is $b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$

$= -\frac{93}{140} = -0.6643$

Coefficient of regression of X on Y is $b_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}$

$= -\frac{93}{398} = -0.2337$

Equation of the line of regression of X on Y is $x - \bar{x} = b_{XY}(y - \bar{y})$

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

$\Rightarrow x = -0.2337y + 0.2337 \times 38 + 32$

$\Rightarrow x = -0.2337y + 40.8806$

Equation of the line of regression of Y on X is $y - \bar{y} = b_{YX}(x - \bar{x})$

$\Rightarrow y - 38 = -0.6642(y - 38)$

$\Rightarrow y = -0.6642x + 0.6642 \times 32 + 38$

$\Rightarrow y = -0.6642x + 59.2576$

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

$$= -0.6643 \times (-0.2337)$$

$$= 0.1552$$

$$r = \pm\sqrt{0.1552}$$

$$= \pm 0.394$$

Now we have to find the most likely marks in statistics (Y) when marks in Economics (X) are 30. We use the line of regression of Y on X.

$\Rightarrow y = -0.6642x + 59.2576$

Put $x = 30$ we get

$\Rightarrow y = -0.6642 \times 30 + 59.2576$

$\Rightarrow y = 39.3286$

2. The two lines of regression are $8x - 10y + 66 = 0, 40x - 18y - 214 = 0$. The variance of X is 9. Find (i) the mean value of X and Y (ii) correlation coefficient between X and Y.

Solution:

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines.

$(1) \times 5 \Rightarrow 40\bar{x} - 50\bar{y} = -330$

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

Subtracting (1) – (2) we get

$\Rightarrow 32\bar{y} = 544$

$\Rightarrow \bar{y} = 17$

Sub $\bar{y} = 17$ in (1) we get,

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

$\Rightarrow 8\bar{x} - 10 \times 17 = -66$

$\Rightarrow 8\bar{x} = -66 + 170$

$\Rightarrow \bar{x} = 13$

Hence the mean value ae given by $\bar{x} = 13$ and $\bar{y} = 17$

(ii) Let us suppose that equation (A) is the equation of line of regression of Y on X and (B) is the equation of the line regression of X on Y, we get after rewriting (A) and (B)

$\Rightarrow 10y = 8x + 66$

$\Rightarrow y = \frac{8}{10}x + \frac{66}{10}$

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

$\Rightarrow 40x = 18y + 214$

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

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

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

$$= \frac{8}{10} \times \frac{18}{40} = \frac{9}{25}$$

$$r = \pm\frac{3}{5}$$

$$= \pm 0.6$$

Since both the regression coefficients are positive, r must be positive.

Hence $r = 0.6$

Important note:

If we take equation (A) as the line of regression of X on Y we get,

$\Rightarrow 8x = 10y - 66$

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

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

$\Rightarrow 18y = 40x - 214$

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

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

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

$$= \frac{10}{8} \times \frac{40}{8} = \frac{25}{9}$$

$$r = 2.78$$

But r^2 should always lies between 0 and 1. Hence our assumption that line (A) is line of regression of X on Y and the line (B) is line of regression of Y on X is wrong.

3. The two lines of regression are $4x - 5y + 33 = 0, 20x - 9y - 107 = 0$. The variance of X is 25. Find (i) the mean value of X and Y (ii) correlation coefficient between X and Y.

Solution:

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines.

$(1) \Rightarrow 20\bar{x} - 9\bar{y} = 107$

$(2) \times 5 \Rightarrow 20\bar{x} - 25\bar{y} = -165$

Subtracting (1) – (2) we get

$\Rightarrow 16\bar{y} = 272$

$\Rightarrow \bar{y} = 17$

Sub $\bar{y} = 17$ in (1) we get,

$(2) \Rightarrow 4\bar{x} - 5\bar{y} = -33$

$\Rightarrow 4\bar{x} - 5 \times 17 = -33$

$\Rightarrow 4\bar{x} = -33 + 85$

$\Rightarrow \bar{x} = 13$

Hence the mean value ae given by $\bar{x} = 13$ and $\bar{y} = 17$

(ii) Let us suppose that equation (A) is the equation of line of regression of Y on X and (B) is the equation of the line regression of X on Y, we get after rewriting (A) and (B)

$\Rightarrow 5y = 4x + 33$

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

$\Rightarrow b_{YX} = \frac{4}{5}$

$\Rightarrow 20x = 9y + 107$

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

$\Rightarrow b_{XY} = \frac{9}{20}$

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

$$= \frac{4}{5} \times \frac{9}{20} = \frac{3}{5}$$

$$r = \pm 0.6$$

4. Can $Y = 5 + 2.8X$ and $X = 3 - 0.5Y$ be the estimated regression equations of Y on X and X on Y respectively? Explain your answer.

Solution:

Given,

$\Rightarrow X = 3 - 0.5Y$

$\Rightarrow b_{XY} = -0.5$

$\Rightarrow Y = 5 + 2.8X$

$\Rightarrow b_{YX} = 2.8$

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

$$= 2.8 \times (-0.5) = -1.4$$

$r = \sqrt{-1.4}$ which is imaginary quantity.

Here r cannot be imaginary.

Hence the given lines are not estimated as regression equations.