$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{4\sigma_X^2}{\sigma_X \sigma_Y} = 4\frac{\sigma_X}{\sigma_Y} \quad \text{(i)}$$

But, $\sigma_Y^2 = E(Y^2) - [E(Y)]^2$

$= E[(4X + 9)^2] - [E(4X + 9)]^2$

$= E(16X^2 + 81 + 72X) - [4E(X) + 9]^2$

$= 16E(X^2) + 81 + 72E(X) - 16[E(X)]^2 - 81 - 72E(X)$

$= 16\{E(X^2) - [E(X)]^2\} = 16\sigma_X^2$

$\Rightarrow \quad \sigma_Y = 4\sigma_X$

Substituting in Eq. (i), we get

$$\rho_{XY} = 4\frac{\sigma_X}{4\sigma_X} = 1$$

EXAMPLE 5.68 X, Y and Z are uncorrelated random variables with 0 mean and SD 5, 12 and 9 respectively. If $U = X + Y$ and $V = Y + Z$, find the correlation coefficient between U and V. **[AU April '03]**

Solution Given: $E(X) = E(Y) = E(Z) = 0$

$E(U) = E(X) + E(Y) = 0$

and $E(V) = E(Y) + E(Z) = 0$

$E(UV) = E[(X + Y)(Y + Z)]$

$= E(XY + XZ + Y^2 + YZ)$

$= E(XY) + E(XZ) + E(Y^2) + E(YZ)$

$\because$ X, Y and Z are uncorrelated

$\text{Cov}(X, Y) = \text{Cov}(Y, Z) = \text{Cov}(X, Z) = 0$

So, $E(XY) = E(X)\,E(Y) = 0$

and $E(XZ) = E(YZ) = 0$

$\because$ $E(UV) = E(Y^2)$ (i)

Given: $\text{Var}(X) = \text{SD}^2 = 5^2 = 25$, $\text{Var}(Y) = 12^2 = 144$,

$\text{Var}(Z) = 9^2 = 81$

$$\rho = \frac{\text{Cov}(U, V)}{\sqrt{\text{Var}(U)\text{Var}(V)}} = \frac{E(UV) - E(U)E(V)}{\sigma_U\,\sigma_V} \quad \text{(ii)}$$

$\text{Var}(X) = E(X^2) - [E(X)]^2 \Rightarrow 5^2 = E(X^2) - 0 \Rightarrow E(X^2) = 25$

Similarly,

$E(Y^2) = 144$

and $E(Z^2) = 81$

$\therefore$ From Eq. (i), $E(UV) = 144$

From Eq. (ii) $\rho = \dfrac{144 - 0}{\sigma_U \cdot \sigma_V} = \dfrac{144}{\sigma_U \cdot \sigma_V}$

$\sigma_U^2 = E(U^2) - [E(U)]^2$

$= E[(X + Y)^2] - 0$

$= E(X^2) + E(Y^2) + 2E(XY)$

$= 25 + 144 = 169$

$\therefore$ $\sigma_U = 13$

$\sigma_V^2 = E[(Y + Z)^2]$

$= E(Y^2) + E(Z^2) + 2E(YZ)$

$= 144 + 81 = 225$

$\therefore$ $\sigma_V = 15$

Substituting in Eq. (ii), we get

$$\rho = \frac{144}{13 \times 15} = \frac{144}{195} = \frac{48}{65}$$

5.5 REGRESSION

The term 'regression' literally means 'stepping back towards the average'.

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

5.5.1 Line of Regression

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of best fit and is obtained by the principle of least squares. The regression equation of Y on X is obtained on minimizing the sum of the squares of the errors parallel to Y axis while the regression equation of X on Y is obtained on minimizing the sum of the squares of the errors parallel to X axis.

If the variables in a bivariate distribution are related, then the points in the scatter diagram (which is a diagrammatic representation of bivariate data) will cluster round some curve called the *curve of regression*. If the curve is a straight line, it is called the *line of regression* and said to be linear regression between the variables, otherwise regression is said to be curvilinear.

5.5.2 Equations of Lines of Regression

The regression line of y on x is

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

which is used to predict or estimate the value of y for any given x.

The coefficient of x in the regression line of y on x is the regression coefficient of y on x and is denoted by

$$b_{yx} = \rho \frac{\sigma_y}{\sigma_x}$$

The regression line of x on y is

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

which is used to predict or estimate the value of x for any given y.

The coefficient of y in the regression line of x on y is the regression coefficient of x on y and is denoted by

$$b_{xy} = \rho \frac{\sigma_x}{\sigma_y}$$

$\therefore$ The correlation coefficient ρ in terms of the regression coefficients is

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

5.5.3 Properties of Correlation and Regression Coefficients

1. Correlation coefficient is the geometric mean between regression coefficients, i.e.
$$\rho = \pm\sqrt{b_{xy}\, b_{yx}}\,.$$
2. The correlation coefficient cannot numerically exceed the arithmetic mean between regression coefficient.
3. If one of the regression coefficients is greater than unity, the other must be less than unity numerically.
4. Regression coefficients are independent of the origin, but not scale.
5. The correlation coefficient and the two regression coefficients have the same sign.

The regression coefficients are obtained by the following expressions in case of discrete values of X and Y:

$$b_{yx} = \rho \frac{\sigma_y}{\sigma_x} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

and

$$b_{xy} = \rho \frac{\sigma_x}{\sigma_y} = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma y^2 - (\Sigma y)^2} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

5.5.4 Angle between the Regression Lines

The angle between the regression lines is given by

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

Note:

(i) When $r = \pm 1$, $\tan\theta = 0$, i.e. $\theta = 0$ or π. In this case, the two lines of regression coincide. When $\theta = \frac{\pi}{2}$, the lines of regression are perpendicular.

(ii) Whenever the two lines intersect, there are two angles between them, one acute angle and the other obtuse angle.

(iii) Both regression lines pass through the point $(\bar{x}, \bar{y})$ where $\bar{x}$ = mean of x and $\bar{y}$ = mean of y.

(iv) Regression curve of Y on X is $y = E(Y/X = x)$ and regression curve of X on Y is $x = E(X/Y = y)$.

EXAMPLE 5.69 Calculate the correlation coefficient for the following heights (in inches) of fathers X and their sons Y. **[AU June '06, November '07]**

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

Solution **Method 1:**

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
$\Sigma(X) = 544$	$\Sigma(Y) = 552$	$\Sigma XY = 37560$	$\Sigma X^2 = 37028$	$\Sigma Y^2 = 38132$

Now, $\bar{X} = \frac{544}{8} = 68$

$\bar{Y} = \frac{552}{8} = 69$

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

$\sigma_X = \sqrt{\frac{1}{n}\Sigma X^2 - \bar{X}^2} = \sqrt{\frac{37028}{8} - 4624} = 2.121$

$\sigma_Y = \sqrt{\frac{1}{n}\Sigma Y^2 - \bar{Y}^2} = \sqrt{\frac{38132}{8} - 4761} = 2.345$

$$r(X, Y) = \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sigma_X \sigma_Y} = \frac{\frac{1}{8} \times 37560 - 4692}{2.121 \times 2.345} = 0.6030$$

Note: Correlation coefficient is independent of change of origin and scale, i.e.

$$r(X, Y) = r(U, V)$$

where

$$U = \frac{X - a}{h}$$

and

$$V = \frac{Y - b}{K}$$

Here a and b are some arbitrary constants and usually the mid-values of the given data X and Y respectively.

Method 2:

X	Y	$U = X - 68$	$V = Y - 68$	UV	U^2	V^2
65	67	−3	−1	3	9	1
66	68	−2	0	0	4	0
67	65	−1	−3	3	1	9
67	68	−1	0	0	1	0
68	72	0	4	0	0	16
69	72	1	4	4	1	16
70	69	2	1	2	4	1
72	71	4	3	12	16	9
		$\Sigma U = 0$	$\Sigma V = 8$	$\Sigma UV = 24$	$\Sigma U^2 = 36$	$\Sigma V^2 = 52$

Now, $\bar{U} = \frac{\Sigma U}{n} = \frac{0}{8} = 0$

$\bar{V} = \frac{\Sigma V}{n} = \frac{8}{8} = 1$

$\text{Cov}(X, Y) = \text{Cov}(U, V)$

$= \frac{\Sigma UV}{n} - \bar{U}\bar{V} = \frac{24}{8} - 0 = 3$

$\sigma_U = \sqrt{\frac{1}{n}\Sigma U^2 - \bar{U}^2} = \sqrt{\frac{36}{8} - 0} = 2.121$

$\sigma_V = \sqrt{\frac{1}{n}\Sigma V^2 - \bar{V}^2} = \sqrt{\frac{52}{8} - 1} = 2.345$

$r(X, Y) = r(U, V)$

$= \frac{\text{Cov}(U, V)}{\sigma_U \cdot \sigma_V} = \frac{3}{2.121 \times 2.345}$

$= 0.6031$

EXAMPLE 5.70 Find the coefficient of correlation between industrial production and export using the following data: **[AU May '06]**

Production (x)	55	56	58	59	60	60	62
Export (y)	35	38	37	39	44	43	44

Solution

x	y	$X = x - 58$	$Y = y - 40$	XY	X^2	Y^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
		$\Sigma X = 4$	$\Sigma Y = 0$	$\Sigma XY = 48$	$\Sigma X^2 = 38$	$\Sigma Y^2 = 80$

$\bar{X} = \frac{\Sigma X}{n} = \frac{4}{7} = 0.5714$

$\bar{Y} = \frac{\Sigma Y}{n} = \frac{0}{7} = 0$

$\text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \bar{X}\bar{Y} = \frac{48}{7} - 0 = 6.857$

$$\sigma_X = \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} = \sqrt{\frac{38}{7} - (0.5714)^2} = 2.2588$$

$$\sigma_Y = \sqrt{\frac{\Sigma Y^2}{n} - \bar{X}\bar{Y}} = \sqrt{\frac{80}{7} - 0} = 3.38$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{6.857}{2.258 \times 3.38} = 0.898$$

EXAMPLE 5.71 Find the correlation coefficient for the following data:

x	10	14	18	22	26	30
y	18	12	24	6	30	36

Solution

x	y	$X = \frac{x-22}{4}$	$Y = \frac{y-24}{6}$	XY	X^2	Y^2
10	18	−3	−1	3	9	1
14	12	−2	−2	4	4	4
18	24	−1	0	0	1	0
22	6	0	−3	0	0	9
26	30	1	1	1	1	1
30	36	2	2	4	4	4
		$\Sigma X = -3$	$\Sigma Y = -3$	$\Sigma XY = 12$	$\Sigma X^2 = 19$	$\Sigma Y^2 = 19$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{-3}{6} = -0.5$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{-3}{6} = -0.5$$

$$\text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \bar{X}\bar{Y}$$

$$= \frac{12}{6} - (-0.5)(-0.5) = \frac{7}{4}$$

$$\sigma_X = \sqrt{\frac{\Sigma X^2}{n} - \bar{X}^2} = \sqrt{\frac{19}{6} - (-0.5)^2} = 1.708$$

$$\sigma_Y = \sqrt{\frac{\Sigma Y^2}{n} - \bar{X}\bar{Y}} = \sqrt{\frac{19}{6} - (-0.5)^2} = 1.708$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{7/4}{1.708 \times 1.708} = 0.5999$$

EXAMPLE 5.72 Calculate the correlation coefficient and the lines of regression from the following data: **[AU December '04, June '06]**

x	62	64	65	69	70	71	72	74
y	126	125	139	145	165	152	180	208

Solution

x	y	$X = x - 70$	$Y = y - 165$	XY	X^2	Y^2
62	126	−8	−39	312	64	1521
64	125	−6	−40	240	36	1600
65	139	−5	−26	130	25	676
69	145	−1	−20	20	1	400
70	165	0	0	0	0	0
71	152	1	−13	−13	1	169
72	180	2	15	30	4	225
74	208	4	43	172	16	1849
		$\Sigma X = -13$	$\Sigma Y = -80$	$\Sigma XY = 891$	$\Sigma X^2 = 147$	$\Sigma Y^2 = 6440$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{8} \times (-13) = \frac{-13}{8}$$

$$\bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times (-80) = \frac{-80}{8} = -10$$

$$\frac{1}{n}\Sigma XY = \frac{1}{8} \times (891) = 111.375$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{8} \times (147) - \left(\frac{-13}{8}\right)^2 = 15.73$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{8} \times (6440) - (-10)^2 = 705$$

$$\rho_{XY} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\bar{Y}}{\sigma_X \sigma_Y} = \frac{111.375 - \left(\frac{-13}{8}\right)(-10)}{3.96 \times 26.55} \Rightarrow \rho_{XY} = 0.9$$

To obtain the regression lines:
The regression line of x on y is

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

which is used to predict or estimate the value of x for any given y.

$$\bar{X} = \frac{-13}{8}$$

$$\bar{Y} = -10$$

$$\bar{x} = \bar{X} + 70 = \frac{-13}{8} + 70 = 68.375$$

$$\bar{y} = \bar{Y} + 165 = -10 + 165 = 155$$

$$\Rightarrow \quad x - 68.375 = 0.9 \times \frac{3.96}{26.55}(y - 155)$$

$$x - 68.375 = 0.134(y - 155)$$

$$x = 0.134y - 20.8 + 68.375$$

$$\Rightarrow \quad x = 0.134y + 47.5$$

The regression line of y on x is

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

$$\Rightarrow \quad y - 155 = 0.9 \times \frac{26.55}{3.96}(x - 68.375)$$

$$y = 6.034x - 412.58 + 155$$

$$\Rightarrow \quad y = 6.034x - 257.58$$

EXAMPLE 5.73 Calculate the correlation coefficient between the variables x and y from the following data:

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Solution We know that the correlation coefficient of X and Y is

$$\rho_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}$$

where

$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{n}\Sigma y^2 - \bar{y}^2$$

x	y	xy	x^2	y^2
1	9	9	1	81
2	8	16	4	64
3	10	30	9	100
4	12	48	16	144
5	11	55	25	121
6	13	78	36	169
7	14	98	49	196
$\Sigma x = 28$	$\Sigma y = 77$	$\Sigma xy = 334$	$\Sigma x^2 = 140$	$\Sigma y^2 = 875$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{1}{7} \times 28 = 4$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{1}{7} \times 77 = 11$$

$$\frac{1}{n}\Sigma xy = \frac{1}{7} \times 334 = \frac{334}{7}$$

$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - \bar{x}^2 = \frac{1}{7} \times 140 - 4^2 = 4$$

$$\sigma_y^2 = \frac{1}{n}\Sigma y^2 - \bar{y}^2 = \frac{1}{7} \times 875 - 11^2 = 4$$

$$\rho_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{\frac{334}{7} - (11 \times 4)}{\sqrt{4}\sqrt{4}} = \frac{13}{14} = 0.928$$

EXAMPLE 5.74 Calculate the correlation coefficient and the lines of regression from the following data:

x	22	26	29	30	31	31	34	35
y	20	20	21	29	27	24	27	31

Solution Using the given data,

x	y	xy	x^2	y^2
22	20	440	484	400
26	20	520	676	400
29	21	609	841	441
30	29	870	900	841
31	27	837	961	729
31	24	744	961	576
34	27	918	1156	729
35	31	1085	1225	961
$\Sigma x = 238$	$\Sigma y = 199$	$\Sigma xy = 6023$	$\Sigma x^2 = 7204$	$\Sigma y^2 = 5077$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{1}{8} \times 238 = 29.75$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{1}{8} \times 199 = 24.875$$

$$\frac{1}{n}\Sigma xy = \frac{1}{8} \times 6023 = 752.875$$

$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - \bar{x}^2 = \frac{1}{8} \times 7204 - 29.75^2 = 15.4375$$

$$\sigma_y^2 = \frac{1}{n}\Sigma y^2 - \bar{y}^2 = \frac{1}{8}\times 5077 - 24.875^2 = 15.859$$

$$\rho_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{752.875 - 740.00}{\sqrt{15.4375 \times 15.859}} = 0.82$$

To obtain the regression lines of x on y:

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

$$(x - 29.75) = 0.82 \times \sqrt{\frac{15.4375}{15.859}}(y - 24.875)$$

$$x - 29.75 = 0.81y - 20.125$$

$$x = 0.81y + 9.63$$

The regression line of y on x:

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

$$(y - 24.875) = 0.82 \times \sqrt{\frac{15.859}{15.4375}}(x - 29.75)$$

$$y - 24.875 = 0.83x - 24.73$$

$$y = 0.83x + 0.15$$

EXAMPLE 5.75 Calculate the correlation coefficient, lines of regression and angle between them from the following data:

x	4	7	5	6	8	5	6	6	4	9
y	2.5	6	4.5	5	4.5	2	3	4.5	3	5.5

Solution Using the given data,

x	y	xy	x^2	y^2
4	2.5	10	16	6.25
7	6	42	49	36
5	4.5	22.5	25	20.25
6	5	30	36	25
8	4.5	36	64	20.25
5	2	10	25	4
6	3	18	36	9
6	4.5	27	36	20.25
4	3	12	16	9
9	5.5	49.5	81	30.25
$\Sigma x = 60$	$\Sigma y = 40.5$	$\Sigma xy = 257$	$\Sigma x^2 = 384$	$\Sigma y^2 = 180.25$

$$\bar{x} = \frac{1}{n}\Sigma x = \frac{60}{10} = 6$$

$$\bar{y} = \frac{1}{n}\Sigma y = \frac{40.5}{10} = 4.05$$

$$\frac{\Sigma xy}{n} = \frac{257}{10} = 25.7$$

$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - \bar{x}^2 = \frac{384}{10} - 6^2 = 2.4$$

$$\sigma_y^2 = \frac{1}{n}\Sigma y^2 - \bar{y}^2 = \frac{180.25}{10} - 4.05^2 = 1.6225$$

Regression lines are

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

$$x - 6 = 0.709 \times \sqrt{\frac{2.4}{1.6225}}(y - 4.05)$$

$$x = 2.505 + 0.86y$$

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

$$y - 4.05 = 0.709 \times \sqrt{\frac{1.6}{2.4}}(x - 6)$$

$$y = 0.58x + 0.55$$

We know that

$$\rho_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sigma_x \sigma_y} = \frac{25.7 - 6 \times 4.05}{\sqrt{2.4 \times 1.6225}} = 0.709$$

$\therefore$ The correlation coefficient $\rho_{xy} = 0.709$

The angle between the regression lines is

$$\tan\theta = \frac{1-\rho^2}{\rho}\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} = \frac{1-(0.709)^2}{0.709} \times \frac{2.4 \times 1.6225}{(2.4)^2 + (1.6225)^2} = 0.32545$$

$\therefore$ $\theta = 0.31464$

EXAMPLE 5.76 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.

[AU June '06, May '07, December '09]

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	$x-\bar{x}$ $= x-32$	$y-\bar{y}$ $= y-38$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
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	–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
Σx $= 320$	Σy $= 380$	$\Sigma(x-\bar{x})$ $= 0$	$\Sigma(y-\bar{y})$ $= 0$	$\Sigma(x-\bar{x})^2$ $= 140$	$\Sigma(y-\bar{y})^2$ $= 398$	$\Sigma(x-\bar{x})(y-\bar{y})$ $= -93$

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

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

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

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 32 = -0.2337(y - 38)$$
$$x = -0.2337y + 40.8806$$

Equation of the line of regression of y on x is

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

$$y - 38 = -0.6643(x - 32)$$
$$y = -0.6643x + 59.2576$$

Coefficient of correlation,

$$\rho^2 = b_{xy} \times b_{yx}$$
$$= -0.6643 \times (-0.2337) = 0.1552$$
$$\rho = \pm 0.394$$

Now, to find the most likely marks in statistics (y) when marks in economics (x) is 30. We use the line of regression of y on x, i.e.

$$y = -0.6643x + 59.2576$$

Put $x = 30$, we get

$$y = -0.6643 \times 30 + 59.2576 = 39.3286$$

EXAMPLE 5.77 Heights of fathers and sons are given in centimetres.

Heights of fathers (x)	150	152	155	157	160	161	164	166
Heights of sons (y)	154	156	158	159	160	162	161	164

Find the two lines of regression and calculate the expected average height of the son when the height of the father is 154 cm.

[AU December '07]

Solution Let 160 and 159 be assumed means of x and y.
Using the given data, we get the following table:

x	y	X $= x - 160$	Y $= y - 159$	X^2	Y^2	XY
150	154	–10	–5	100	25	50
152	156	–8	–3	64	9	24
155	158	–5	–1	25	1	5
157	159	–3	0	9	0	0
160	160	0	1	0	1	0
161	162	1	3	1	9	3
164	161	4	2	16	4	8
166	164	6	5	36	25	30
		$\Sigma X = -15$	$\Sigma Y = 2$	$\Sigma X^2 = 251$	$\Sigma Y^2 = 74$	$\Sigma XY = 120$

$$\bar{x} = 160 + \frac{\Sigma X}{n} = 160 - \frac{15}{8} = 158.13$$

$$\bar{y} = 159 + \frac{\Sigma Y}{n} = 159 + \frac{2}{8} = 159.25$$

Since regression coefficients are independent of change of origin, we have regression coefficient of y on x.

$$b_{yx} = b_{YX} = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{8\times 120 - (-15)\times 2}{8\times 251 - (-15)^2} = 0.555$$

Regression coefficient of x on y is

$$b_{xy} = b_{XY} = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2} = \frac{990}{8\times 74 - 4} = 1.68$$

Equation of the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$
$$x - 158.13 = 1.68(y - 159.25)$$
$$x = 1.68y - 109.41$$

Equation of the line regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$
$$y - 159.25 = 0.56(x - 158.13)$$
$$y = 0.56x + 70.697$$

When $x = 154$

$$y = 0.56(154) + 70.697$$
$$= 156.937$$

Note:

$$b_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{X}\bar{Y}}{\frac{1}{n}\Sigma x^2 - \bar{X}^2} = \frac{\frac{1}{n}\Sigma xy - \frac{\Sigma X}{n}\frac{\Sigma Y}{n}}{\frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma X}{n}\right)^2} = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

EXAMPLE 5.78 If $\bar{x} = 970, \bar{y} = 18, \sigma_x = 38, \sigma_y = 2$ and $\rho = 0.6$, find the line of regression of x on y. **[AU December '09]**

Solution Given: $\bar{x} = 970,\ \bar{y} = 18,\ \sigma_x = 38,\ \sigma_y = 2,\ \rho = 0.6$
The regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}) = \rho\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

$$\therefore \quad (x - 970) = 0.6\left(\frac{38}{2}\right)(y - 18)$$
$$x = 970 + (0.6)\,(19)\,(y - 18)$$
$$= 764.8 + 11.4y$$

$\therefore$ The regression line of x on y is

$$x = 11.4y + 764.8$$

EXAMPLE 5.79 A statistical investigation obtains the following regression equation in a survey:

$$x - y = 6$$
and
$$0.64x + 4.08 = 0$$

Here x denotes the age of husband and y denotes the age of wife. Find

(i) the means of x and y,
(ii) the correlation coefficient between x and y and σ_y = SD of y, if σ_x = SD of x = 4.

Solution By one of the properties of regression lines, all the regression lines pass through their means ($\bar{x}$ and $\bar{y}$).
$\therefore$ $(\bar{x}, \bar{y})$ satisfies the equation.

(i) $\bar{x} - \bar{y} = 6$

and $0.64\,\bar{x} + 4.08 = 0 \Rightarrow \bar{x} = \frac{-4.08}{0.64} = -6.37$

$-6.37 - \bar{y} = 6 \Rightarrow \bar{y} = -12.37$

(ii) Given: $x - y = 6 \Rightarrow y = x - 6$

and $x = 0y - 6.37$

The regression coefficient of x on y is the coefficient of y in the regression line of x on y, i.e.

$$\frac{\rho\sigma_x}{\sigma_y} = \text{coefficient of } y = b_{xy} = 0$$

$$\frac{\rho\sigma_y}{\sigma_x} = \text{coefficient of } x = b_{yx} = 1$$

$$\rho^2 = b_{xy}b_{yx} = 0(1) = 0 \Rightarrow \rho = 0$$

Given: $\sigma_x = 4$

But, $$\rho\frac{\sigma_x}{\sigma_y} = 0 \Rightarrow \rho\sigma_x = 0\times\sigma_y$$
$$0\times 4 = 0\times\sigma_y$$

$\Rightarrow$ σ_y can assume any value.

EXAMPLE 5.80 The two regression lines are $4x - 5y + 33 = 0$ and $20x - 9y = 107$, and Var(X) = 25. Find

(i) the means of x and y,
(ii) the values of ρ and σ_y, and
(iii) the angle between the regression lines.

[AU June '06, December '07]

Solution

(i) Given: $4x - 5y = -33$
$20x - 9y = 107$

Since all the regression lines pass through the means,

$$4\bar{x} - 5\bar{y} = -33 \quad \text{(i)}$$
$$20\bar{x} - 9\bar{y} = 107 \quad \text{(ii)}$$

Multiplying Eq. (i) by 5,

$$20\bar{x} - 25\bar{y} = -165 \quad \text{(iii)}$$

Equation (iii)–Eq. (ii) gives

$$-16\bar{y} = -272$$
$$\bar{y} = 17$$

Substituting in Eq. (i), we get

$$4\bar{x} = -33 + 5(17)$$
$$\bar{x} = \frac{85 - 33}{4} = 13$$

(ii) $20x - 9y = 107$

$$x = \frac{9}{20}y + \frac{107}{20}$$
$$x = 0.45y + 5.35$$

$$b_{xy} = \frac{\rho\sigma_x}{\sigma_y} = 0.45 = \text{coefficient of } y$$

$$5y = 33 + 4x$$
$$y = \frac{4}{5}x + \frac{33}{5}$$
$$= 0.8x + 6.6$$

$$b_{yx} = \frac{\rho\sigma_y}{\sigma_x} = 0.8 = \text{coefficient of } x$$
$$\rho^2 = b_{xy} \times b_{yx} = 0.36$$

$\therefore$ The correlation coefficient $\rho = 0.6$

(iii) Given: $\sigma_x^2 = 25$

$\therefore$ $\sigma_x = 5$

$$b_{xy} = \frac{\rho\sigma_x}{\sigma_y} = 0.45 \Rightarrow \sigma_y = \frac{0.6 \times 5}{0.45}$$
$$= 6.67$$

The angle between the regression lines are

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

$$\Rightarrow \quad \tan\theta = \frac{1-(0.6)^2}{0.6} \times \left(\frac{5 \times 6.67}{25 + 44.49}\right)$$

$$\theta = \tan^{-1}(0.512) \Rightarrow \theta = 27.11$$

EXAMPLE 5.81 If X, Y denote the deviation of variance from the arithmetic mean and if $\rho = 0.5$, $\Sigma XY = 120$, $\sigma_y = 8$, $\Sigma X^2 = 90$. Find n, number of times.

Solution Given $X = x - \bar{x}, Y = y - \bar{y}$

$$\therefore \quad \rho = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}} = \frac{\frac{1}{n}\Sigma XY}{\sqrt{\frac{1}{n}\Sigma X^2}\sqrt{\sigma_y^2}}$$

$$0.5 = \frac{\frac{1}{n}(120)}{\frac{1}{\sqrt{n}} \times 9.48 \times 8}$$

$$\sqrt{n} = \frac{120}{9.48 \times 8 \times 0.5} = 3.16$$

$$n = 10$$

EXAMPLE 5.82 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. **[AU June '06, November '07]**

Solution Given: $X = 3 - 0.5Y \Rightarrow b_{XY} = -0.5$

$Y = 5 + 2.8X \Rightarrow b_{YX} = 2.8$

$\therefore$ $\rho^2 = b_{XY} \times b_{YX} = (-0.5) \times (2.8) = -1.4$

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

ρ cannot be imaginary.

$\therefore$ The given lines are not estimated as regression equations.

EXAMPLE 5.83 If $y = 2x - 3$ and $y = 5x + 7$ are the two regression lines, find the mean values of x and y. Find the correlation coefficient between x and y. Find an estimate of x when $y = 1$.

Solution The two regression lines always pass through their means $(\bar{x}, \bar{y})$

$\therefore$ We have

$$\bar{y} = 2\bar{x} - 3$$

$$\bar{y} = 5\bar{x} + 7 \Rightarrow 2\bar{x} - \bar{y} = 3$$
$$5\bar{x} - \bar{y} = -7$$

Solving we get

$$\bar{x} = \frac{-10}{3}$$
$$\bar{y} = \frac{-29}{3}$$

From the given equations, we have

$$x = \frac{1}{5}y - \frac{7}{5} \Rightarrow b_{xy} = \frac{1}{5}$$
$$y = 2x - 3 \Rightarrow b_{yx} = 2$$

$\therefore$ The correlation coefficient $\rho_{xy} = b_{xy}\, b_{yx} = \frac{2}{5}$

When $y = 1$, $x = \frac{1}{5} - \frac{7}{5} = \frac{-6}{5}$

Aliter

If we choose the regression equation for x as $x = \frac{1}{2}y + \frac{3}{2}$ and for y as $y = 5x + 7$, then $b_{xy} = \frac{1}{2}, b_{yx} = 5$

$\therefore$ $$\rho = \frac{1}{2} \times 5 = 2.5 > 1$$

$\therefore$ It is not possible.

EXAMPLE 5.84 Given that $x = 4y + 5$ and $y = kx + 4$ are the lines of regression of x on y and y on x respectively, show that $0 < 4k < 1$. If $k = 1/16$, find the means of two variables and coefficient of correlation between them.

Solution Given: $x = 4y + 5$, regression line of x on y

$\therefore$ $b_{xy} = 4$

The regression line of y on x is

$$y = kx + 4$$

$\therefore$ $$b_{yx} = k$$
$$\rho^2 = b_{xy}\, b_{yx} = 4k$$

But, $0 \le \rho^2 \le 1 \Rightarrow 0 \le 4k \le 1$

If $k = \frac{1}{16}$, then

$$\rho^2 = 4k = \frac{4}{16} = \frac{1}{4} \Rightarrow \rho = \pm\frac{1}{2}$$

Since b_{xy} and b_{yx} are positive, ρ is also positive.

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

The two lines of regression are

$$x = 4y + 5$$
$$y = \frac{x}{16} + 4$$

Since the regression lines pass through their means

$$\bar{x} = 4\bar{y} + 5 \quad \text{(i)}$$
$$\bar{y} = \frac{\bar{x}}{16} + 4 \quad \text{(ii)}$$

Equation (i) + 4 × Eq. (ii) gives

$$\bar{x} = \frac{\bar{x}}{4} + 21 \Rightarrow 3\bar{x} = 84 \Rightarrow \bar{x} = 28$$

Substituting in Eq. (i)

$$28 = 4\bar{y} + 5 \Rightarrow \bar{y} = 5.75$$

$\therefore$ $$\rho = \frac{1}{2}$$
$$\bar{x} = 28$$
$$\bar{y} = 5.75$$

EXAMPLE 5.85 In a bivariate population $\sigma_x = \sigma_y = \sigma$ and the angle between the regression lines is $\tan^{-1}(6)$, obtain the value of correlation coefficient.

Solution We know that the angle between the regression lines is

$$\tan\theta = \frac{1-\rho^2}{\rho}\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}$$
$$= \frac{1-\rho^2}{\rho}\frac{\sigma^2}{2\sigma^2}$$
$$\tan\theta = \frac{1-\rho^2}{2\rho}$$
$$\theta = \tan^{-1}\left(\frac{1-\rho^2}{2\rho}\right) = \tan^{-1}(6) \text{ (given)}$$

$\Rightarrow$ $$\frac{1-\rho^2}{2\rho} = 6$$