Correlation

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

Positive | Direct Correlation

If the increase | decrease in one results in the corresponding increase | decrease in the other, then the variables are positively correlated.

Ex: Correlation between (i) heights and weights of a group of people (ii) Income and expenditure.

Negative | Diverse Correlation

If the increase | decrease in one results in corresponding decrease | increase in other, then the variables are negatively correlated.

Ex: Correlation between demand and price.

Karl Pearson's Coefficient of Correlation

Correlation Coefficient between two random variables X and Y, denoted by $r(X,Y)$ or r_{XY} is a numerical measure of linear relationship between them and is defined as

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

where $Cov(X,Y) = \sigma_{XY} = \frac{1}{n} \Sigma (x_i - \bar{x})(y_i - \bar{y})$

or $\sigma_{XY} = \frac{1}{n} \Sigma x_i y_i - \bar{x}\bar{y}$,

$$\sigma_X^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 \text{ or } \sigma_X^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2,$$

$$\sigma_Y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2 \text{ or } \sigma_Y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2.$$

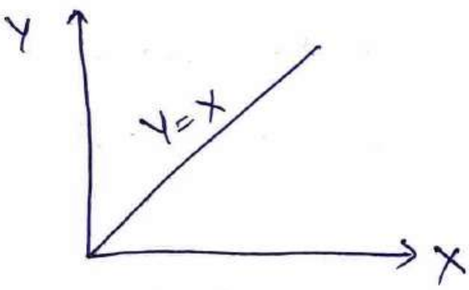
Range of Correlation Coefficient

$$-1 \le r \le 1.$$

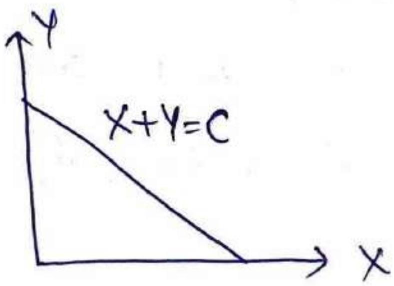
If $r = -1$, then correlation is perfect and negative.

If $r = 1$, then correlation is perfect and positive.

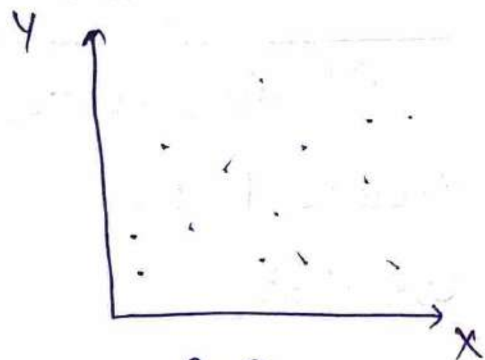
Scatter Plots | diagrams



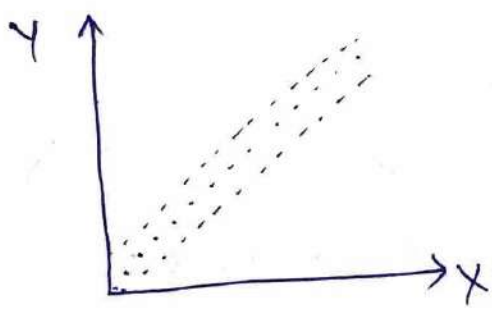
$r = +1$



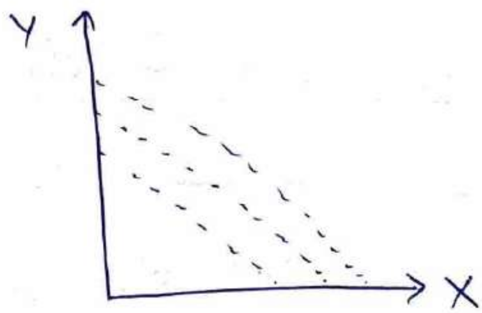
$r = -1$



$r = 0$



$r > 0$



$r < 0$

Results 1. Correlation Coefficient is independent of change of origin and scale.

2. If X and Y are random variables and a, b, c, d are any numbers provided only that $a \neq 0$, $c \neq 0$, then

$$r(aX+b, cY+d) = \frac{ac}{|ac|} r(X,Y).$$

3. Two independent variables are uncorrelated.

Ex Calculate the correlation coefficient for the following ~~heigth~~ heights (in inches) of fathers (X) and their sons (Y):

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

Sol: Let $U = X - 68$, $V = Y - 69$

X	Y	U	V	U^2	V^2	UV
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
		0	0	36	44	24

$\bar{U} = 0$, $\bar{V} = 0$, $\sigma_{UV} = \frac{1}{n} \sum u_i v_i - \bar{u}\bar{v} = \frac{1}{8}(24) = 3$

$\sigma_U^2 = \frac{1}{n} \sum u_i^2 - \bar{u}^2 = \frac{1}{8}(36) = 4.5$

$\sigma_V^2 = \frac{1}{n} \sum v_i^2 - \bar{v}^2 = \frac{1}{8}(44) = 5.5$

$\therefore\ r(U,V) = \frac{\sigma_{UV}}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4.5 \times 5.5}} = \frac{3}{\sqrt{24.75}} = \frac{3}{4.97}$

$\Rightarrow\ r(U,V) = 0.604 = r(X,Y)$

$\boxed{r(X,Y) = 0.604}$

Ex A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results:

$n = 25,\ \Sigma X = 125,\ \Sigma X^2 = 650,\ \Sigma Y = 100,\ \Sigma Y^2 = 460,\ \Sigma XY = 508.$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

Sol: Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 36 - 64 + 64 + 36 = 650$

Corrected $\Sigma Y^2 = 460 - 196 - 36 + 144 + 64 = 436$

Corrected $\Sigma XY = 508 - 84 - 48 + \cancel{84}\ 96 + 48 = 520$

$\overline{X} = \frac{\Sigma X}{n} = \frac{125}{25} = 5,\quad \overline{Y} = \frac{1}{n}\Sigma Y = \frac{100}{25} = 4$

$\sigma_{XY} = \frac{1}{n}\Sigma XY - \overline{X}\,\overline{Y} = \frac{1}{25}(520) - 20 = 0.8$

$\sigma_x^2 = \frac{1}{n} \Sigma x^2 - \bar{x}^2 = \frac{1}{25}(650) - 25 = 1$

$\sigma_y^2 = \frac{1}{n} \Sigma y^2 - \bar{y}^2 = \frac{1}{25}(436) - 16 = 1.44$

$\therefore$ Corrected $r(x,y) = \frac{0.8}{\sqrt{1 \times 1.44}} = \frac{0.8}{1.2} = 0.67.$

$\boxed{r(x,y) = 0.67}$

Correlation coefficient for bi-variate data and probability distribution

$g(x) = \sum_{y} f(x,y)$ is the sum of frequencies along any column and $h(y) = \sum_{x} f(x,y)$ is the sum of frequencies along any row.

Here $\sum_{x}\sum_{y} f(x,y) = \sum_{y}\sum_{x} f(x,y) = \sum_{x} g(x) = \sum_{y} h(y) = N$

$$\bar{x} = \frac{1}{N}\sum_{x}\sum_{y} x\, f(x,y) = \frac{1}{N}\sum_{x} x\, g(x),$$

$$\bar{y} = \frac{1}{N}\sum_{y} y\, h(y), \quad \sigma_x^2 = \frac{1}{N}\sum_{x} x^2 g(x) - \bar{x}^2,$$

$$\sigma_y^2 = \frac{1}{N}\sum_{y} y^2 h(y) - \bar{y}^2.$$

Ex The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.

Ages → / Marks ↓	18	19	20	21	Total
10-20	4	2	2	—	8
20-30	5	4	6	4	19
30-40	6	8	10	11	35
40-50	4	4	6	8	22
50-60	–	2	4	4	10
60-70	—	2	3	1	6
Total	19	22	31	28	100

Calculate the correlation coefficient.

Sol: Let $U = X - 19$, $V = \frac{Y - 35}{10}$

Correlation table

V	Mid Value	U Age X → Marks Y ↓	-1 18	0 19	1 20	2 21	$f(v)$	$v f(v)$	$v^2 f(v)$	$\sum_u uv f(u,v)$
-2	15	10-20	4 (8)	2 (0)	2 (-4)	—	8	-16	32	4
-1	25	20-30	5 (5)	4 (0)	6 (-6)	4 (-8)	19	-19	19	-9
0	35	30-40	6 (0)	8 (0)	18 (0)	11 (0)	35	0	0	0
1	45	40-50	4 (-4)	4 (0)	6 (6)	8 (16)	22	22	22	18
2	55	50-60	—	2 (0)	4 (8)	4 (16)	10	20	40	24
3	65	60-70	—	2 (0)	3 (9)	1 (6)	6	18	54	15
		$g(u)$	19	22	31	28	100	25	167	52
		$u g(u)$	-19	0	31	56	68			
		$u^2 g(u)$	19	0	31	112	162			
		$\sum_v uv f(u,v)$	9	0	13	30	52			

$$\bar{U} = \frac{1}{N} \sum_u U g(U) = \frac{68}{100} = 0.68,$$

$$\bar{V} = \frac{1}{N} \sum_v V f(V) = \frac{25}{100} = 0.25$$

$$\sigma_U^2 = \frac{1}{N} \sum U^2 g(U) - \bar{U}^2 = \frac{162}{100} - (0.68)^2 = 1.1576$$

$\cancel{\sigma_v^2 = \frac{1}{N}\sum_v v^2 g(v) =}$

$$\sigma_V^2 = \frac{1}{N}\sum_v v^2 h(v) - \bar{v}^2 = \frac{167}{100} - (0.25)^2 = 1.6075$$

$$\sigma_{UV} = \frac{1}{N}\sum_u \sum_v uv\, f(u,v) - \bar{u}\bar{v} = \frac{52}{100} - 0.68 \times 0.25$$
$$= 0.35$$

$$\therefore r(U,V) = \frac{0.35}{\sqrt{1.6075 \times 1.1576}} = \frac{0.35}{1.36} = 0.26$$

Since, correlation coefficient is independent of change of origin and scale,

$$r(X,Y) = r(U,V) = 0.26.$$

<u>Ex</u> The joint pdf of X and Y is given below:

Y \ X	-1	1	$h(y)$
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
$g(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

Find the correlation coefficient between X and Y.

<u>Sol</u>: $r(X,Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$,

where $\sigma_{XY} = \frac{1}{N}\sum_x \sum_y xy\, f(x,y) - \bar{x}\bar{y}$

$\sigma_X^2 = \frac{1}{N}\sum_x x^2 g(x) - \bar{x}^2$, $\sigma_Y^2 = \frac{1}{N}\sum_y y^2 h(y) - \bar{y}^2$

$$\bar{x} = \frac{1}{N}\sum_x x g(x) = \frac{1}{2}\left[-\frac{3}{8} + \frac{5}{8}\right] = \frac{2}{8} = \frac{1}{4}$$

$$\bar{x} = \frac{1}{4}, \quad \bar{y} = \frac{1}{N}\sum_y y h(y) = \frac{4}{8} = \frac{1}{2}.$$

$$\sigma_{xy} = 0 - \frac{2}{8} + 0 + \frac{2}{8} = 0 - \frac{1}{8} = \frac{-1}{8}$$

$$\sigma_x^2 = \frac{3}{8} + \frac{5}{8} - \frac{1}{16} = \frac{15}{16},$$

$$\sigma_y^2 = 0 + \frac{4}{8} - \frac{1}{4} = \frac{2}{8} = \frac{1}{4}$$

$$\therefore \; r(x,y) = \frac{\frac{-1}{8}}{\sqrt{\frac{15}{16} \times \frac{1}{4}}} = \frac{-0.125}{0.48} = -0.26.$$

$$r(x,y) = -0.26$$

Rank Correlation

Spearman's Rank Correlation Coefficient

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

Ex The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics:

(1,1), (2,10), (3,3), (4,4), (5,5), (6,7), (7,2), (8,6), (9,8), (10,11), (11,15), (12,9), (13,14), (14,12), (15,16), (16,13).

Calculate the rank correlation coefficient for expertise of this group in Mathematics and Physics.

Sol:

Ranks in Maths (X)	Ranks in Physics (Y)	$d = X - Y$	d^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

(X)	(Y)	$d = X - Y$	d^2
14	12	2	4
15	16	-1	1
16	13	3	9
			136

$$\therefore \rho = 1 - \frac{6(136)}{16(256-1)} = 1 - \frac{816}{4080} = 1 - 0.2 = 0.8$$

$$\boxed{\rho = 0.8}$$

Ex Ten competitors in a musical test were ranked by the three judges A, B and C in the following order:

Ranks by A :	1	6	5	10	3	2	4	9	7	8
Ranks by B :	3	5	8	4	7	10	2	1	6	9
Ranks by C :	6	4	9	8	1	2	3	10	5	7

Discuss which pair of judges has the nearest approach to common likings in music.

Sol:

X	Y	Z	$d_1 = X - Y$	$d_2 = X - Z$	$d_3 = Y - Z$	d_1^2	d_2^2	d_3^2
1	3	6	-2	-5	-3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	$-\cancel{4}3$	-4	-1	$\cancel{16}9$	16	1
10	4	8	$\cancel{2}6$	2	-4	$\cancel{4}36$	4	16
3	7	1	$\cancel{2}-4$	2	6	$\cancel{4}16$	4	36
2	10	2	$\cancel{8}-8$	0	8	64	0	64
4	2	3	2	1	-1	4	1	1
9	1	10	8	-1	-9	64	1	81
7	6	5	1	$\cancel{3}2$	1	1	$\cancel{36}4$	1
8	9	7	-1	1	2	1	1	4
						200	60	214

$$\rho(X,Y) = 1 - 6\frac{\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(200)}{10(99)} = 1 - \frac{1200}{990} = 1 - 1.21$$
$$= -0.21$$

$\rho(X,Y) = -0.21$

$$\rho(X,Z) = 1 - \frac{6(60)}{990} = 1 - \frac{360}{990} = 1 - 0.36 = 0.64$$

$$\rho(Y,Z) = 1 - \frac{6(214)}{990} = 1 - 1.297 = -0.297$$

Since, $\rho(X,Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

Tied Ranks | Repeated Ranks

$$\rho = 1 - \frac{6(\Sigma d^2 + \text{Correction factor})}{n(n^2-1)}$$

To Σd^2, we add $\frac{m(m^2-1)}{12}$ for each repeated value.

Ex: Obtain the rank correlation coefficient for the following data:

X :	68	64	75	50	64	80	75	40	55	64
Y :	62	58	68	45	81	60	68	48	50	70.

Sol :

X	Y	(x) Rank X	(y) Rank Y	d = x - y	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
				$\Sigma d = 0$	72

$$\rho = 1 - \frac{6\left(\Sigma d^2 + 2 + \frac{6}{12} + \frac{6}{12}\right)}{10(99)}$$

(64) $\frac{m(m^2-1)}{12} = \frac{3(9-1)}{12} = \frac{24}{12} = 2$

(75) $\frac{2(4-1)}{12} = \frac{6}{12}$

(68) $\frac{6}{12}$

$$= 1 - \frac{6\left(72 + 2 + \frac{1}{2} + \frac{1}{2}\right)}{990}$$

$$= 1 - \frac{450}{990} = 1 - 0.45 = 0.55$$

$$\boxed{\rho = 0.55}$$

Remark Limits for Rank Correlation coefficient:

$$-1 \leq \rho \leq 1.$$

Remarks 1. $\Sigma d = 0$

2. Karl Pearson's correlation coefficient assumes that the parent population from which sample observations are drawn is normal, whereas Spearman's correlation coefficient is non-parametric.

Regression

Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.

In regression, one variable is considered as an independent variable and another variable is taken as dependent variable.

Regression is used to predict the value of dependent variable on the basis of the value of the independent variable.

Linear regression

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram 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 there is said to be linear regression between the variables, otherwise regression is said to be curvilinear.

As the line of regression gives the best estimate of the value of one variable for any specific value of the other variable, it is called line of "best fit".

Line of Regression of Y on X is

$Y = a + bX$ $\quad$ b→ slope of line, $\dfrac{\text{change in y co-ordinate w.r.t}}{\text{change in x co-ordinate}}$

and is given by

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

Where $b_{YX} = r\dfrac{\sigma_Y}{\sigma_X}$ is the regression coefficient of Y on X.

Line of Regression of X on Y is

$X = a + bY$

and is given by

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

where $b_{XY} = r\dfrac{\sigma_X}{\sigma_Y}$ is the regression coefficient of X on Y.

Ex Obtain the equations of two lines of regression for the following data. Also obtain the estimate of X for Y = 70.

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

Sol: Let $U = X - 68$, $V = Y - 69$.

$\bar{U} = 0$, $\bar{V} = 0$, $\sigma_U^2 = 4.5$, $\sigma_V^2 = 5.5$, $\sigma_{UV} = 3$,

$r(U,V) = 0.604$

Since, correlation coefficient is independent of change of origin, $r(X,Y) = r(U,V) = 0.604$

Now, $\bar{U} = \bar{X} - 68$, $\bar{V} = \bar{Y} - 69$

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

Since, variance is independent of change of origin,

$\therefore \sigma_X^2 = \sigma_U^2 = 4.5$, $\sigma_Y^2 = \sigma_V^2 = 4.5$

$\Rightarrow \sigma_X = \sqrt{4.5} = 2.12$, $\sigma_Y = 2.34$

Equation of line of regression of Y on X is

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

where $b_{YX} = \dfrac{r\sigma_Y}{\sigma_X} = (0.604)\dfrac{(2.34)}{2.12} = 0.67$

$\therefore \quad Y - 69 = 0.67(X - 68)$

$\Rightarrow \quad Y = 0.67X - 45.33 + 69$

$\Rightarrow \quad \boxed{Y = 0.67X + 23.66}$

Equation of line of regression of X on Y is

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

where $b_{XY} = \dfrac{r\sigma_X}{\sigma_Y} = (0.604)\dfrac{2.12}{2.34} = 0.55$

$\therefore \quad X - 68 = 0.55(Y - 69)$

$\Rightarrow \quad X = 0.55Y - 37.95 + 68$

$\Rightarrow \quad \boxed{X = 0.55Y + 30.05}$

If Y is 70, estimated value of X is

$$\hat{X} = 0.55(Y) + 30.05 = 0.55(70) + 30.05$$

$\hat{X} = 68.55$, where $\hat{X}$ is estimate of X.

Properties of Regression coefficients

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

$$r^2 = b_{xy} \times b_{yx} \Rightarrow r = \pm\sqrt{b_{xy} \times b_{yx}}$$

* b_{xy}, b_{yx}, r all have same sign.

(2) If one of the regression coefficient is greater than 1 then other must be less than one. $[\because r^2 \leq 1]$

(3) The modulus value of the arithmetic mean of regression coefficients is not less than the modulus value of the correlation coefficient r.

$$\left|\frac{b_{xy} + b_{yx}}{2}\right| > |r|.$$

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

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

$$b_{xy} = \frac{h}{k} b_{uv} \text{ and } b_{yx} = \frac{k}{h} b_{vu}$$

(5) The point of intersection of both regression lines is $(\bar{X}, \bar{Y})$.

Ex In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are readable:

Variance of $X = 9$. Regression equations: $8X - 10Y + 66 = 0$, $40X - 18Y = 214$. What are (i) the mean values X and Y, (ii) the correlation coefficient between X and Y, and (iii) the standard deviation of Y?

Sol: (i) Since, both lines pass through $(\bar{X}, \bar{Y})$.

$$\therefore \quad 8\bar{X} - 10\bar{Y} + 66 = 0 \Rightarrow 4\bar{X} - 5\bar{Y} = -33 \quad -①$$

$$40\bar{X} - 18\bar{Y} = 214. \Rightarrow 20\bar{X} - 9\bar{Y} = 107 \quad -②$$

$$① \times 5 \Rightarrow 20\bar{X} - 25\bar{Y} = -165 \quad -③$$

② − ③ provides, $16\bar{Y} = 272$

$$\Rightarrow \bar{Y} = 17$$

and $4\bar{X} = -33 + 85 = 52$

$$\Rightarrow \bar{X} = 13$$

$\therefore (\bar{X}, \bar{Y})$ is $(13, 17)$.

So, $\bar{X} = 13$, $\bar{Y} = 17$.

(ii) Let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the lines of regression of Y on X and X on Y, respectively.

$$\therefore \quad Y = \frac{8}{10}X + \frac{66}{10} \text{ and } X = \frac{18}{40}Y + \frac{214}{40}$$

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

$$\Rightarrow b_{YX} = \frac{4}{5}, \; b_{XY} = \frac{9}{20}$$

X on Y

$8X - 10Y + 66 = 0$

$$X = \frac{10Y}{8} - \frac{66}{8}$$

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

Y on X

$40X - 18Y = 214$

$$\Rightarrow Y = \frac{40X}{18} - \frac{214}{18}$$

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

$\therefore r^2 = \frac{4}{5} \cdot \frac{9}{20} = 0.36$

$\Rightarrow r = \pm 0.6$

$r^2 = b_{xy} \, b_{yx} = \frac{10}{8} \cdot \frac{40}{18}$

$r^2 = 2.78$

But, $0 < r^2 < 1$.

Since, both b_{xy} and b_{yx} are positive,

$r = 0.6$.

(iii) $b_{xy} = r \frac{\sigma_x}{\sigma_y} \Rightarrow \frac{9}{20} = (0.6) \frac{3}{\sigma_y} \Rightarrow \sigma_y = \frac{1.8 \times 20}{9}$

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

or $b_{yx} = r \frac{\sigma_y}{\sigma_x} \Rightarrow \frac{8}{10} = (0.6) \frac{\sigma_y}{3} \Rightarrow \sigma_y = \frac{8}{10} \times \frac{3}{0.6}$

$\boxed{\sigma_y = 4}$

Ex Find the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata from the following:

	Kolkata	Mumbai
Average price	65	67
Standard deviation	2.5	3.5

Correlation coefficient between the prices of commodities in the two cities is 0.8.

Sol: Let the prices in Kolkata and Mumbai be denoted by X and Y respectively. We have,

$\bar{X} = 65, \bar{Y} = 67, \sigma_X = 2.5, \sigma_Y = 3.5, r(X,Y) = 0.8$

Line of regression of Y on X is

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

$\Rightarrow Y - 67 = 1.12X - 72.8$

$\Rightarrow Y = 1.12X - 5.8$

When, $X = 70$, $\hat{Y} = 1.12(70) - 5.8 = 72.6$

Hence, the most likely price in Mumbai corresponding to the price of Rs. 70 at Kolkata is Rs. 72.6.

Ex Can $Y = 5 + 2.8X$ and $X = 3 - 0.5Y$ be the estimated regression equations of Y on X and X on Y respectively?

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

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

This is not possible as the signs of b_{XY} and b_{YX} should be same.

Angle between two lines of Regression

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

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

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

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

In this case two lines of regression either coincide

or they are parallel to each other.
But, since both lines of regression pass through the point $(\bar{x}, \bar{y})$, they cannot be parallel.
Hence, in the case of perfect correlation, positive or negative, two lines of regression coincide.

Fitting of a curve

(a) Fitting a straight line :- $Y = a + bX$

$$\left.\begin{aligned} \Sigma Y &= na + b\Sigma X \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}\right\}$$ Normal equations

(b) Fitting a parabola :- $Y = a + bX + cX^2$

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma x^2 y + a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

Properties of Correlation Coefficient

1. $-1 \leq r \leq 1$ or $|Cov(X,Y)| \leq \sigma_X \cdot \sigma_Y$

Proof :- $r(X,Y) = \dfrac{Cov(X,Y)}{\sigma_X \sigma_Y} = \dfrac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\left[\frac{1}{n}\sum(x_i-\bar{x})^2 \frac{1}{n}\sum(y_i-\bar{y})^2\right]^{1/2}}$

$\Rightarrow r^2(X,Y) = \dfrac{\left(\sum a_i b_i\right)^2}{\sum a_i^2 \sum b_i^2}$, where $a_i = x_i - \bar{x}$, $b_i = y_i - \bar{y}$

By Schwartz inequality,

$\left(\sum a_i b_i\right)^2 \leq \sum a_i^2 \sum b_i^2$

$\therefore r^2(X,Y) \leq 1 \Rightarrow |r(X,Y)| \leq 1 \Rightarrow -1 \leq r(X,Y) \leq 1.$

$\Rightarrow \left|\dfrac{Cov(X,Y)}{\sigma_X \sigma_Y}\right| \leq 1$

$\Rightarrow |Cov(X,Y)| \leq \sigma_X \sigma_Y$

2. Correlation Coefficient is independent of change of origin and change of scale.

Proof : Let $U = \dfrac{X-a}{h}$, $V = \dfrac{Y-b}{k}$.

$\therefore X = a + hU$, $Y = b + kV$, where a, b, h, k are constants;

$h > 0, k > 0$.

T.P : $r(X,Y) = r(U,V)$.

Since, $X = a + hU$, $Y = b + kV$.

Taking expectations on both sides, we get,

$$E(X) = a + hE(U), \quad E(Y) = b + kE(V).$$

$$\therefore \quad X - E(X) = h[U - E(U)]$$

$$Y - E(Y) = k[V - E(V)]$$

$$\Rightarrow \quad Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

$$= hk\, E[(U - E(U))(V - E(V))] = hk\,\sigma_{UV}$$

$$\sigma_X^2 = E[(X - E(X))^2] = h^2 E[U - E(U)]^2$$

$$\sigma_Y^2 = E[(Y - E(Y))^2] = k^2 E[(V - E(V))^2]$$

$$\Rightarrow \quad \sigma_X^2 = h^2 E[U - E(U)]^2 \text{ and } \sigma_Y^2 = k^2 E[V - E(V)]^2$$

$$\sigma_X = h\,\sigma_U \text{ and } \sigma_Y = k\,\sigma_V.$$

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \cancel{\frac{E[(U - E(U))(V - E(V))]}{E[U - E(U)]\, E[V - E(V)]}}$$

$$= \frac{\sigma_{UV}}{\sigma_U \sigma_V} = \cancel{Cov(X,Y)}.$$

$$= r(U,V).$$

$$\therefore \quad r(X,Y) = r(U,V).$$

3. Two independent variables are uncorrelated.

<u>Proof</u>:- If X and Y are independent variables, then

$$Cov(X,Y) = 0 \Rightarrow r(X,Y) = 0.$$

Hence two independent variables are uncorrelated.

But, the converse may not be true; i.e., two uncorrelated variables may not be independent.

For ex →

X	-3	-2	-1	1	2	3	$\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	-27	-8	-1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0, \quad \mathrm{Cov}(X,Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = 0.$$

$\therefore\ r(X,Y) = 0.$

Thus, in the above example X and Y are uncorrelated. But, X and Y are not independent but related by the relation $Y = X^2$.

Properties of Regression Coefficients

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

<u>Proof</u>: $b_{xy} = r\dfrac{\sigma_x}{\sigma_y}$ and $b_{yx} = r\dfrac{\sigma_y}{r_x}$

$$\therefore\ b_{xy} \times b_{yx} = r^2 \Rightarrow r = \pm\sqrt{b_{xy} \times b_{yx}}$$

<u>Remark</u>: The sign of correlation coefficient is the same as that of regression coefficients.

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

<u>Proof</u>: Let one of the regression coefficients, say, b_{yx} be greater than unity.

$$b_{yx} > 1 \Rightarrow \frac{1}{b_{yx}} < 1$$

Also, $r^2 \le 1 \Rightarrow b_{xy} \cdot b_{yx} \le 1$

Hence, $b_{xy} \le \dfrac{1}{b_{yx}} < 1$.

(3) The modulus value of the ~~regression~~ arithmetic mean of the regression coefficients is not less than the modulus value of the correlation coefficient r.

<u>Proof</u>: <u>T.P.</u>: $\left|\dfrac{b_{xy} \times b_{yx}}{2}\right| \ge |r|$

$\Rightarrow \left|\frac{1}{2}\left(r\frac{\sigma_X}{\sigma_Y} + r\frac{\sigma_Y}{\sigma_X}\right)\right| \geq |r|$

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

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

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

$\Rightarrow \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y \geq 0$

$\Rightarrow (\sigma_X - \sigma_Y)^2 \geq 0$, which is always true,

since the square of a real quantity ≥ 0.

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

Proof. Let $U = \frac{X-a}{h}$, $V = \frac{Y-b}{k}$, where $a, b, h(>0)$ and $k(>0)$ are constants.

$\Rightarrow X = a + hU,\ Y = b + kV.$

$Cov(X,Y) = hk\ Cov(U,V)$

$\sigma_X^2 = h^2\sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{hk\,\sigma_{UV}}{h^2\sigma_U^2} = \frac{k}{h}\frac{\sigma_{UV}}{\sigma_U^2}$ $\cancel{= \frac{k}{h}\frac{\sigma_{XY}}{hk\sigma_U^2}}$

$b_{VU} = r\frac{\sigma_V}{\sigma_U} = \frac{\sigma_{UV}}{\sigma_U\sigma_V}\cdot\frac{\sigma_V}{\sigma_U} = \frac{\sigma_{UV}}{\sigma_U^2}$

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

Similarly, $b_{XY} = \frac{h}{k} b_{UV}$.

(5) Point of intersection of both regression lines is $(\bar{X}, \bar{Y})$.