CHAPTER

2 Statistical Techniques

1. Introduction

In Statistics so far we have studied problems involving a single variable. Many a time, we come across problems which involve two or more than two variables. Data relating to two variates are called **Bivariate Data**. If we study bivariate data carefully, we may find some relation between the two variables. For example, if a car-owner maintains the record of petrol consumption and mileage, he will find that there is some relation between the two variables. On the other hand, if we compare the figures of rainfall with the production of cars, we may find that there is no relation between the two variables. If there is any relation between two variables *i.e.* if as one variable changes the other also changes in the same or opposite direction, we say that they are correlated. Thus, correlation means **"the study of existence and the magnitude and direction of variation between two or more variables"**.

2. Types of Correlation

Correlation may be classified as :

(1) Positive and negative,

(2) Linear and non-linear.

(1) Positive and Negative Correlation

The distinction between the positive and negative correlation depends upon the direction of change of two variables. If both the variables change in the same direction *i.e.* if, as one variable increases, the other also increases and as one variable decreases, the other also decreases, the correlation is called positive (*e.g.* advertising and sales). If, on the other hand, the variables change in opposite direction *i.e.* if, as one variable increases, the other decreases and vice-versa, then the correlation is called negative (*e.g.* T.V. registrations and cinema attendance).

(2) Linear or Non-linear Correlation

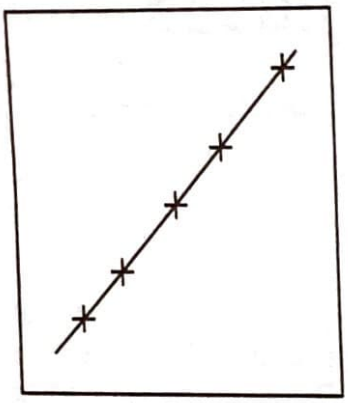
This distinction is based upon the nature of the graph of the relation between the variables. If the graph is a straight line the correlation is called linear and if the graph is not a straight line but a curve it is called non-linear or curvi-linear correlation.

We shall consider the following commonly used methods of studying correlation. (1) Scatter Diagram, (2) Karl Pearson's Coefficient of Correlation, (3) Spearman's Rank-Correlation Coefficient.

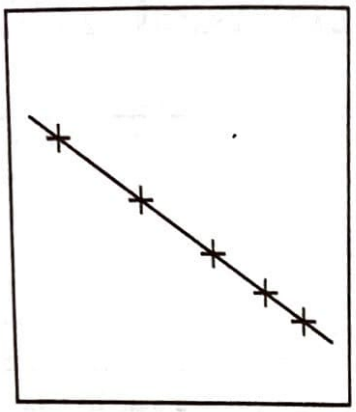
3. Scatter Diagram

One of the most simple methods of studying correlation between two variables is to construct a scatter diagram.

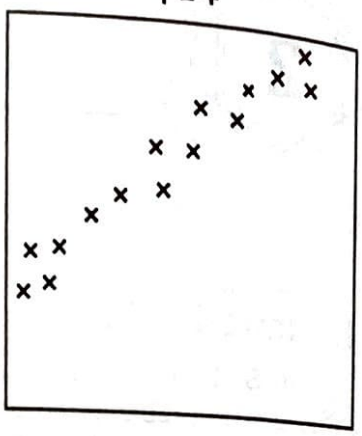
Perfect Positive Correlation
r = +1



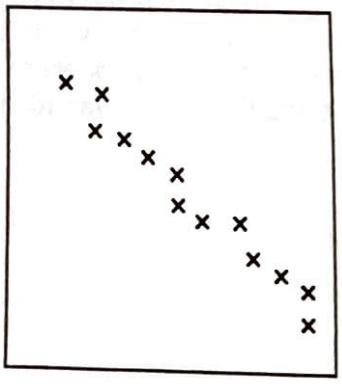
Perfect Negative Correlation
r = –1



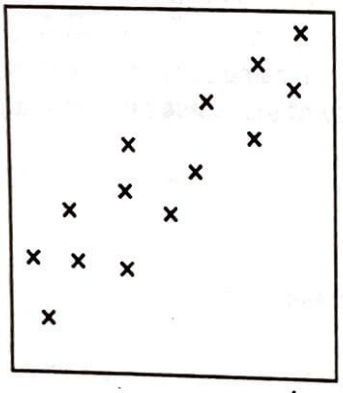
High Degree of Positive Correlation
r = +



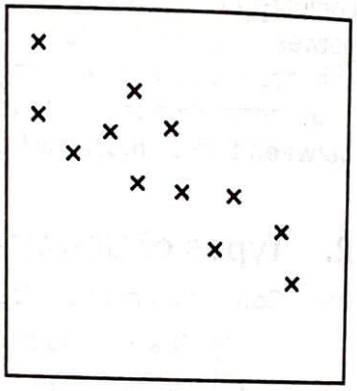
High Degree of Negative Correlation
r = –



Low Degree of Positive Correlation
r = +

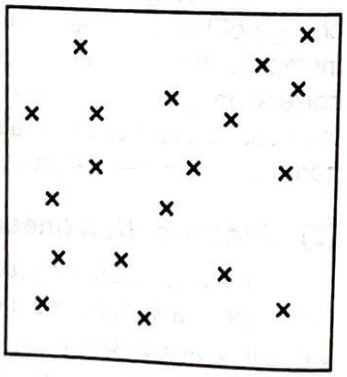


Low Degree of Negative Correlation
r = –



To obtain a scatter diagram, one variable is plotted along the *x*-axis and the other along the *y*-axis, on a graph paper. By plotting data in this way, we get points which are generally scattered but which show a pattern. The way in which the points are scattered indicates the degree and direction of correlation. If the points are close to each other we infer that the variables are correlated. If they are spread away from each other, we infer that the variables are not correlated. Moreover, if the points lie in a narrow strip rising from left-hand bottom to the right-hand top, we say that there is positive correlation of high order. If the points lie in a narrow strip, falling from the left-hand top to the right-hand bottom, we say that there is negative correlation of high order. If the points are all spread over, we say that there is zero correlation.

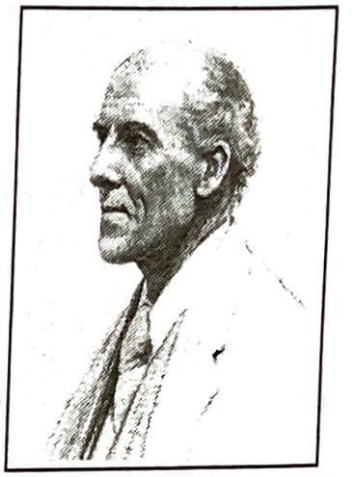
No Correlation
r = 0



4. Karl Pearson's Coefficient of Correlation

The method of scatter diagram is descriptive in nature and gives only a general idea of correlation. The most commonly used method which gives a mathematical expression for correlation is the one suggested by Karl Pearson (1867-1936) a British Biometrician.

Karl Pearson (1857 - 1936)



Born in London, he went to King's College, Cambridge in 1876 to study mathematics graduating in 1879 as Third Wrangler in the Mathematical Tripos. He then went to Germany to study Physics at the University of Heidelberg. Other subjects he studied in Germany include metaphysics, physiology, Roman Law, German Literature and Socialism. He studied so many subjects because he believed that there was no subject in the universe unworthy of study. Then he returned to London to study Law, although he never practised. In 1881 he returned to mathematics and was appointed as professor of mathematics at University College, London. In 1891 he met Walter Weldon, a zoologist and worked with him in biometry and evolutionary theory. He was introduced to Galton, Darwins cousin and became Galtons statistical heir. He was the first holder of the Galton's Chair of Eugenics. In 1911 he founded the world's first university statistics department at University College, London. He remained with the department until his retirement in 1933 and continued to work until his death in 1936. He thus established the new discipline of mathematical statistics.

His famous book **"The Grammar of Science"** covers several themes that were later to become part of the theories of Einstein and other scientists. He speculated that an observer who travelled at the speed of light would see an eternal now and an observer who travelled faster than light would see time reversal. He also discussed antimatter, fourth dimension and wrinkles in time.

Karl Pearson was awarded many medals including The Darwin Medal, a DSc from university of London. His commitment to socialism and his ideals led him to refuse the honours of being an OBE (Officer of the Order of the British Empire) and knighthood in 1935.

Karl Pearson is known for Karl Pearson's coefficient of correlation, methods of moments, Pearson's system of continuous curves, Chi-distance, Statistical hypothesis testing theory, Statistical decision theory, Pearson's chi-square test, etc.

Just as $\sigma_x^2 = \frac{1}{N}\Sigma(x-\bar{x})^2$ gives us a measure of variation in x and $\sigma_y^2 = \frac{1}{N}\Sigma(y-\bar{y})^2$ gives a measure of variation in y we may expect $\frac{1}{N}\Sigma(x-\bar{x})(y-\bar{y})$ to give the measure of simultaneous variation in x and y. But this will depend upon the units of x and y. To find a ratio which is independent of these units, we divide it by the quantities of the same order that is by $\sigma_x \cdot \sigma_y$. With this view in mind Karl Pearson suggested in 1890 the following coefficient of correlation to measure correlation between x and y. It is denoted by r.

Thus,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{N\sigma_x\sigma_y} \quad \text{..... (1)}$$

But $\frac{1}{N}\Sigma(x-\bar{x})(y-\bar{y})$ is called the **covariance** between x and y. Hence, from (1), we have

$$r = \frac{\text{cov.}(x,y)}{\sigma_x \cdot \sigma_y} \quad \text{..... (2)}$$

If we put $\sigma_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$, $\sigma_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{N}}$, then

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}} \quad \ldots\ldots (3)$$

If we write $x - \bar{x} = x'$, $y - \bar{y} = y'$, then

$$r = \frac{\Sigma x'y'}{\sqrt{\Sigma x'^2 \cdot \Sigma y'^2}} \quad \ldots\ldots (4)$$

The Karl Pearson's coefficient of correlation is also called the **product moment coefficient correlation.**

Further, we can expand (3) and write

$$r = \frac{\Sigma(xy - x\bar{y} - \bar{x}y + \bar{x}\bar{y})}{\sqrt{\Sigma(x^2 - 2x\bar{x} + \bar{x}^2) \cdot \Sigma(y^2 - 2y\bar{y} + \bar{y}^2)}}$$

$$= \frac{\Sigma xy - \bar{y}\Sigma x - \bar{x}\Sigma y + \bar{x}\bar{y}\Sigma 1}{\sqrt{(\Sigma x^2 - 2\bar{x}\Sigma x + \bar{x}^2\Sigma 1) \cdot (\Sigma y^2 - 2\bar{y}\Sigma y + \bar{y}^2\Sigma 1)}}$$

But $\Sigma x = N\bar{x}$, $\Sigma y = N\bar{y}$ and $\Sigma 1 = N$

$$\therefore \quad r = \frac{\Sigma xy - N\bar{x}\bar{y}}{\sqrt{(\Sigma x^2 - N\bar{x}^2) \cdot (\Sigma y^2 - N\bar{y}^2)}} \quad \ldots\ldots (5)$$

If $\bar{x}$, $\bar{y}$ are integers we take deviations of x and y from them and use the formula (3). If we have to find r from direct values we use the formula (5). This is the most *commonly used formula.*

(i) Limits for r : $-1 \le r \le 1$

Proof : If we write $E(X) = \mu_X$, $E(Y) = \mu_Y$, then **(M.U. 2004)**

$$E\left[\left(\frac{X-\mu_X}{\sigma_X}\right) \pm \left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]^2 \ge 0$$

$$\therefore \quad E\left(\frac{X-\mu_X}{\sigma_X}\right)^2 + E\left(\frac{Y-\mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{E\left[(X-\mu_X)^2(Y-\mu_Y)\right]}{\sigma_X \Sigma_Y} \ge 0$$

$\therefore \quad 1 + 1 \pm 2r \ge 0 \qquad \therefore \quad 1 \pm r \ge 0$

$\therefore \quad 1 + r \ge 0 \qquad$ or $\qquad 1 - r \ge 0$

$\therefore \quad 1 \ge -r \qquad$ or $\qquad 1 \ge r$

$\therefore \quad -1 \le r,\ r \le 1 \qquad \therefore \quad -1 \le r \le 1 \qquad [\because E(X-\mu_X)^2 = \sigma_X^2]$

(ii) Theorems on correlation

Theorem 1 : If *x, y* are independent variables they are not correlated.

We accept this theorem without proof.

Theorem 2 : Correlation coefficient is independent of change of origin and change of scale. **(M.U. 2002, 07)**

This means if we write $u_i = \frac{x-a}{h}$, $v_i = \frac{y-b}{k}$, then

$$r_{xy} = r_{uv} \qquad \text{..... (6)}$$

i.e., the correlation between x and y is equal to the correlation between u and v.

We accept this theorem without proof.

Remark ....

The above theorem can also be stated as :-

"If $x = au + b$, $y = cv + d$ where a, b, c, d are constants then $r_{xy} = r_{uv}$."

Example : Discuss the statement : "If the coefficient of correlation between x and y is negative then the coefficient of correlation between $(-x)$ and $(-y)$ is positive. **(M.U. 1998)**

Sol. : If we write $a = 0$ and $b = 0$, $h = -1$ and $k = -1$ in (A), then by the above theorem since $r_{xy} = r_{uv}$, the coefficient of correlation between $-x$ and $-y$ will be also the same in magnitude and sign as the coefficient of correlation between x and y.

Or since the coefficient of correlation does not change under change of scale and since $-x$ and $-y$ mean the change of scale, the coefficient of correlation between $-x$ and $-y$ will be also negative.

Theorem 3 : *If d_x and d_y denote the deviations of x and y from the assumed means A and B then*

$$r = \frac{\Sigma d_x d_y - \dfrac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \dfrac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \dfrac{(\Sigma d_y)^2}{N}}} \qquad \text{..... (7)}$$

We accept this result without proof.

5. Interpretation of the Coefficient of Correlation

1. $r > 0{\cdot}95$: If r is greater than 0·95, it indicates high degree of correlation and the value of one variable can be estimated from a known value of the other fairly accurately.

2. $r > 0{\cdot}75$ but $< 0{\cdot}85$: If r is greater than 0·75 but less than 0·85, there is probably a definite relationship between the variables and the value of one variable can be roughly estimated from a known value of the other.

3. $r > 0{\cdot}40$ but $< 0{\cdot}60$: If r is greater than 0·40 but less than 0·60 there may be some relationship between the two variables. But the value of one variable calculated from a known value of the other cannot be reliable.

4. $r < 0{\cdot}35$: If r is less than 0·35 the correlation is poor and one variable cannot be estimated from the other.

5. r nearly zero : If r is nearly equal to zero, it indicates that there is probably no relation between the two variables *i.e.* they are independent of each other.

6. Computation of Coefficient of Correlation : *(Ungrouped Data)*

There are three methods of calculating r.

(1) Actu[illegible]n method, (2) Step deviation method,

(3) A[illegible]n method.

(1) Actual Mean Method

The formula to be used is, $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

Steps :

(i) Calculate mean $\bar{X}$ and then take deviation x of X from $\bar{X}$ *i.e.*, calculate $x = X - \bar{X}$.

(ii) Calculate mean $\bar{Y}$ and then take deviation y of Y from $\bar{Y}$ *i.e.*, calculate $y = Y - \bar{Y}$.

(iii) Multiply x by y and prepare the column of xy.

(iv) Take the squares of x and prepare the column of x^2.

(v) Take the squares of y and prepare the column of y^2.

(vi) Apply the above formula.

Example 1 : Find from the following values of the demand and the corresponding price of a commodity, the degree of correlation between the demand and price by computing Karl Pearson's coefficient of correlation.

Demand in quintals : 65, 66, 67, 67, 68, 69, 70, 72.

Price in Paise per k.g. : 67, 68, 65, 68, 72, 72, 69, 71.

Sol. : Let X denote the demand in Quintals and Y denote the price in paise per kg.

Calculation of r between demand and price

Sr. No.	Demand in Qnt. $X - \bar{X}$, $\bar{X} = 68$			Price per Kg. $Y - \bar{Y}$, $\bar{Y} = 69$			Product
	X	x	x^2	Y	y	y^2	xy
1	65	− 3	9	67	− 2	4	6
2	66	− 2	4	68	− 1	1	2
3	67	− 1	1	65	− 4	16	4
4	67	− 1	1	68	− 1	1	1
5	68	0	0	72	+ 3	9	0
6	69	+ 1	1	72	+ 3	9	3
7	70	+ 2	4	69	0	0	0
8	72	+ 4	16	71	+ 2	4	8
$N = 8$	$\Sigma X = 544$		$\Sigma x^2 = 36$	$\Sigma Y = 552$		$\Sigma y^2 = 44$	$\Sigma xy = + 24$

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

$\therefore \quad r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$

But $\Sigma xy = 24$, $\Sigma x^2 = 36$, $\Sigma y^2 = 44$.

$\therefore \quad r = \frac{24}{\sqrt{36 \times 44}} = \frac{24}{39 \cdot 80} = 0 \cdot 6030.$

Example 2 : Calculate Karl Pearson's coefficient of correlation for the following bivariate series.

X : 28, 45, 40, 38, 35, 33, 40, 32, 36, 33

Y : 23, 34, 33, 34, 30, 26, 28, 31, 36, 35 **(M.U. 2015)**

Calculationof r between X and Y

Sol. :

Sr. No.	$X-\bar{X}$, $\bar{X}=36$			$Y-\bar{Y}$, $\bar{Y}=31$			Product
	X	x	x^2	Y	y	y^2	xy
1	28	−8	64	23	−8	64	+64
2	45	+9	81	34	+3	9	+27
3	40	+4	16	33	+2	4	+8
4	38	+2	4	34	+3	9	+6
5	35	−1	1	30	−1	1	+1
6	33	−3	9	26	−5	25	−15
7	40	+4	16	28	−3	9	−12
8	32	−4	16	31	0	0	0
9	36	0	0	36	+5	25	0
10	33	−3	9	35	+4	16	−12
$N=10$	$\Sigma X=360$		$\Sigma x^2=216$	$\Sigma Y=310$		$\Sigma y^2=162$	$\Sigma xy=+97$

Now, $\bar{X}=\frac{360}{10}=36$ and $\bar{Y}=\frac{310}{10}=31.$

$\therefore\ r=\frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$

But $\Sigma xy=97,\ \Sigma x^2=216,\ \Sigma y^2=162.$

$\therefore\ r=\frac{97}{\sqrt{216\times 162}}=\frac{97}{187.1}=0.5186.$

(2) Step-deviation Method

As in the case of mean and standard deviation, to simplify calculations we can use step-deviation method whenever possible for calculating r. But it should be noted that *the result is not to be multiplied by the constant in the final stage.* The reason is the coefficient of correlation is independent of change of origin and change of scale. (See Theorem 2 of r, page 2-4)

Example 3 : Calculate the co-efficient of correlation from the following data.

X: 100, 200, 300, 400, 500.

Y: 30, 40, 50, 60, 70.

(M.U. 2015)

Sol. :

Calculations of r between X and Y

Sr. No.	$\bar{X}=300$, $X-\bar{X}$		$\frac{X-\bar{X}}{100}$		$\bar{Y}=50$, $Y-\bar{Y}$		$\frac{Y-\bar{Y}}{10}$		Product
	X	$X-300$	x	x^2	Y	$Y-50$	y	y^2	xy
1	100	−200	−2	4	30	−20	−2	4	4
2	200	−100	−1	1	40	−10	−1	1	1
3	300	0	0	0	50	0	0	0	0
4	400	100	1	1	60	10	1	1	1
5	500	200	2	4	70	20	2	4	4
$N=5$	$\Sigma X=1500$		$\Sigma x^2=10$		$\Sigma Y=250$		$\Sigma y^2=10$		$\Sigma xy=10$

Now $\bar{X} = \frac{1500}{5} = 300$ and $\bar{Y} = \frac{250}{5} = 50$ $\quad \therefore \quad r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$

But $\Sigma xy = 10,\ \Sigma x^2 = 10,\ \Sigma y^2 = 10.$ $\quad \therefore \quad r = \frac{10}{\sqrt{10 \times 10}} = \frac{10}{10} = +1.$

(3) Assumed Mean Method

Since in the calculation of r, deviations are to be squared the calculations will be tedious if the means are not integers but data are in integers. In such cases, we take deviations from an assumed mean conveniently chosen. The corresponding formula is

$$r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

where, d_x = deviations of X from an assumed mean, $(X - A)$,
d_y = deviations of Y from an assumed mean, $(Y - B)$,
N = Number of pairs of observations.

Steps :

(i) Assume any mean A for X and calculate deviations d_x of X from A *i.e.*, $d_x = X - A$.
(ii) Assume any mean B for Y and calculate deviations d_y of Y from B *i.e.*, $d_y = Y - B$.
(iii) Take the squares of d_x.
(iv) Take the squares of d_y.
(v) Take the products of d_x and d_y.
(vi) Apply the formula.

Example 4 : Find the co-efficient of correlation for the prices (in Rs.) and sales units.

Price in Rs. : 100, 98, 85, 92, 90, 84, 88, 90, 93, 95.
Sales Units : 500, 610, 700, 630, 670, 800, 800, 750, 700, 690.

Sol. : Let us assume 92 and 670 to be the means of X and Y respectively.

Calculations of r between price and sale

Sr. No.	Price in ₹ $(X - 92)$			Sales Units $(Y - 670)$			Product
	X	d_x	d_x^2	Y	d_y	d_y^2	$d_x d_y$
1	100	+ 8	64	500	– 170	28900	– 1360
2	98	+ 6	36	610	– 60	3600	– 360
3	85	– 7	49	700	+ 30	900	– 210
4	92	0	0	630	– 40	1600	0
5	90	– 2	4	670	0	0	0
6	84	– 8	64	800	+ 130	16900	– 1040
7	88	– 4	16	800	+ 130	16900	– 520
8	90	– 2	4	750	+ 80	6400	– 160
9	93	+ 1	1	700	+ 30	900	+ 30
10	95	+ 3	9	690	+ 20	400	+ 40
$N = 10$	$\Sigma d_x = -5$		$\Sigma d_x^2 = 247$	$\Sigma d_y = 150$		$\Sigma d_y^2 = 76500$	$\Sigma d_x d_y = -3560$

$$\text{Now, } r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

But, $\Sigma d_x d_y = -3560, \quad \Sigma d_x = -5, \quad \Sigma d_y = 150$

$N = 10, \quad \Sigma d_x^2 = 247, \quad \Sigma d_y^2 = 76500.$

$$\therefore r = \frac{-3560 - \frac{(-5)\times(150)}{10}}{\sqrt{247 - \frac{25}{10}}\sqrt{76500 - \frac{22500}{100}}}$$

$$= \frac{-3560 + 75}{\sqrt{247 - 2{\cdot}5}\sqrt{76500 - 2250}}$$

$$= \frac{-3485}{\sqrt{244{\cdot}5}\sqrt{74250}}$$

$$= \frac{3485}{4261} = -0{\cdot}8179$$

Example 5 : Calculate the correlation coefficient from the following data.

X : 23, 27, 28, 29, 30, 31, 33, 35, 36, 39.

Y : 18, 22, 23, 24, 25, 26, 28, 29, 30, 32. **(M.U. 2004, 14, 19)**

Sol. : Let us assume 30 and 25 to be the means of x and y respectively.

Calculation of r between X and Y

Sr. No.	$(X-30)$			$(Y-25)$			Product
	X	d_x	d_x^2	Y	d_y	d_y^2	$d_x d_y$
1	23	−7	49	18	−7	49	49
2	27	−3	9	22	−3	9	9
3	28	−2	4	23	−2	4	4
4	29	−1	1	24	−1	1	1
5	30	0	0	25	0	0	0
6	31	+1	1	26	+1	1	1
7	33	+3	9	28	+3	9	9
8	35	+5	25	29	+4	16	20
9	36	+6	36	30	+5	25	30
10	39	+9	81	32	+7	49	63
$N = 10$		$\Sigma d_x = +11$	$\Sigma d_x^2 = 215$		$\Sigma d_y = +7$	$\Sigma d_y^2 = 163$	$\Sigma d_x d_y = 186$

$$\text{Now, } r = \frac{\Sigma d_x d_y - \frac{(\Sigma d_x)(\Sigma d_y)}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

But, $\Sigma d_x d_y = 186,\ \Sigma d_x = 11,\ \Sigma d_y = 7$

$N = 10,\ \Sigma d_x^2 = 215,\ \Sigma d_y^2 = 163.$

$$\therefore\ r = \frac{186 - \frac{11 \times 7}{10}}{\sqrt{215 - \frac{(11)^2}{10}}\sqrt{163 - \frac{(7)^2}{10}}}$$

$$= \frac{186 - 77}{\sqrt{215 - 12{\cdot}1}\sqrt{163 - 4{\cdot}9}}$$

$$= \frac{178{\cdot}3}{\sqrt{202{\cdot}9}\sqrt{158{\cdot}1}} = 0{\cdot}9948$$

7. Direct Method of Calculating Coefficient of Correlation

We can find the coefficient of correlation directly without taking the deviations of x from their respective means. In such cases the following formula is used.

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

where, x and y are the observed values of the variables and $\bar{x}, \bar{y}$ are their respective me

The formula can also be written as,

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

$$r = \frac{\Sigma xy - n\bar{x}\bar{y}}{n\sigma_x\sigma_y}$$

Example 1 : Calculate the coefficient of correlation between X and Y from the follow

X : 3, 5, 4, 6, 2.

Y : 3, 4, 5, 2, 6.

Sol. :

Calculations of r

x	x^2	y	y^2	xy
3	9	3	9	9
5	25	4	16	20
4	16	5	25	20
6	36	2	4	12
2	4	6	36	12
$\Sigma x = 20$	$\Sigma x^2 = 90$	$\Sigma y = 20$	$\Sigma y^2 = 90$	$\Sigma xy = 73$

Since, $\bar{x} = 4$, $\bar{y} = 4$ putting these values in equation (8).

$$\therefore \quad r = \frac{73 - \frac{20^2}{5}}{\sqrt{90 - \frac{20^2}{5}}\sqrt{90 - \frac{20^2}{5}}} = \frac{73 - 80}{10} = -0.7.$$

EXERCISE - I

1. Calculate the coefficient of correlation from the following data. Is there any marked correlation between the production and price of tea ?

Production in crores (kgs)	:	34, 27, 31, 38, 38, 36, 39, 40.
Price in Rs. per kg	:	3·75, 4·62, 4·25, 4·12, 4·28, 4·32, 4·21, 4·05.

[Ans. : $r = -0.48$ **]**

2. Compute the coefficient of correlation between X and Y from their values given blow.

X : 30, 33, 25, 10, 33, 75, 40, 85, 90, 95.

Y : 68, 65, 80, 85, 70, 30, 55, 18, 15, 10. **(M.U. 2015) [Ans. :** $r = -0.7069$ **]**

3. The following data give the hardness (X) and tensile strength (Y) for some specimens of a material in certain units in a factory. Find the correlation coefficient and interpret your result.

X : 23·3, 17·5, 17·8, 20·7, 18·1, 20·9, 22·9, 20·8.

Y : 4·2, 3·8, 4·6, 3·2, 5·2, 4·7, 4·4, 5·6.

[Ans. : $r = -0.072$. No correlation **]**

4. Calculate the product moment coefficient of correlation between the indices of business activity (X) and employment (Y) from the following data.

X : 100, 102, 108, 111, 115, 116, 118.

Y : 110, 100, 104, 108, 112, 116, 120. **[Ans. :** $r = 0.75$ **]**

5. Find Karl Pearson's coefficient of correlation between X and Y.

X : 10, 12, 14, 15, 16, 17, 18, 10, 14, 15

Y : 17, 16, 15, 12, 10, 9, 8, 15, 13, 12 **[Ans. :** $r = -0.93$ **]**

6. Compute a coefficient of correlation between X and Y.

X : 3, 6, 4, 5, 7

Y : 2, 4, 5, 3, 6 **[Ans. :** $r = 0.7$ **]**

7. Calculate the coefficient of correlation between price and demand by direct method.

Price : 2, 3, 4, 7, 4

Demand: 8, 7, 3, 1, 1 **[Ans. :** -0.81 **]**

8. Calculate the coefficient of correlation between the X and Y by direct method.

X : 8, 8, 7, 5, 6, 2

Y : 3, 4, 10, 13, 22, 8 **(M.U. 2017) [Ans. :** 0·2646 **]**

9. Soil temperature (x) and Germination interval (y) for winter wheat in 12 places are as follows.

x **(in °F)** : 57, 42, 38, 42, 45, 42, 44, 40, 46, 44, 43, 40.

y **(days)** : 10, 26, 41, 29, 27, 27, 19, 18, 19, 31, 29, 33.

Calculate the coefficient of correlation between x and y. **(M.U. 2015) [Ans.**

10. Find the Karl Pearson's coefficient of correlation between X and Y from the following.

X : 51, 54, 56, 59, 65, 60, 70

Y : 38, 44, 33, 36, 33, 23, 13

[Ans. : $r = -0.7977$]

8. Spearman's Rank Correlation

(M.U. 2006)

The method developed by Spearman is simpler than Karl Pearson's method since, it depends upon ranks of the items and actual values of the items are not required. Hence, this can be used to study correlation even when actual values are not known. For instance we can study correlation between intelligence and honesty by this method.

Let x_i, y_i be the ranks in the two characteristics of the i-th member where $i = 1, 2, \ldots, n$. We assume that no two members have the same rank either for x or for y. Thus, x and y take all integral values between 1 and n.

$$\therefore \bar{x} = \frac{1}{2}(1+2+3+\ldots..+n) = \frac{n+1}{2}$$

Similarly, $\bar{y} = \frac{1}{2}(1+2+3+\ldots..+n) = \frac{n+1}{2} \quad \therefore \bar{x} = \bar{y}$

$$\therefore \Sigma(x_i - \bar{x})^2 = \Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \Sigma x_i^2 - 2\bar{x}\,\Sigma x_i + \bar{x}^2\Sigma 1$$

$$= \Sigma x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \Sigma x_i^2 - n\bar{x}^2$$

$$= (1^2 + 2^2 + \ldots.. + n^2) - n\left(\frac{n+1}{2}\right)^2$$

$$\therefore \Sigma(x_i - \bar{x})^2 = \frac{n}{6}(n+1)(2n+1) - \frac{n(n+1)^2}{4} = \frac{1}{12}(n^3 - n)$$

Similarly, $\Sigma(y_i - \bar{y}) = \frac{1}{12}(n^3 - n)$.

If d_i denotes the difference between the ranks of i-th member in the two variables, we have

$$d_i = (x_i - y_i) = (x_i - \bar{x}) - (y_i - \bar{y}) = x_i' - y_i' \text{ (since, } \bar{x}, \bar{y} \text{ are equal)}$$

where, x_i', y_i' denote the deviations of x_i, y_i from their means $\bar{x}$, $\bar{y}$ respectively.

$$\therefore \Sigma d_i^2 = \Sigma x_i'^2 + \Sigma y_i'^2 - 2\Sigma x_i' y_i'$$

$$\therefore \Sigma d_i^2 = \frac{1}{12}(n^3 - n) + \frac{1}{12}(n^3 - n) - 2\Sigma x_i' y_i'$$

$$\therefore \Sigma x_i' y_i' = \frac{1}{2}\left[\frac{n^3 - n}{6} - \Sigma d_i^2\right]$$

But the coefficient of correlation

$$= \frac{\Sigma x_i' y_i'}{\sqrt{\Sigma x_i'^2 \cdot \Sigma y_i'^2}} = \frac{\frac{1}{2}\left[\frac{n^3 - n}{6} - \Sigma d_i^2\right]}{\frac{1}{12}(n^3 - n)} = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

This coefficient is denoted by R

$$\therefore \boxed{R = 1 - \frac{6\Sigma d_i^2}{n^3 - n}.} \quad \ldots\ldots (10)$$

The value of R, as of r, lies between +1 and –1. If $R = +1$, there is perfect positive correlation *i.e.* there is complete agreement in the same direction. If $R = -1$, there is perfect negative correlation *i.e.* there is complete agreement but in opposite direction. Generally, the value of R is neither +1 nor –1 but lies somewhere in between. If $R = 0$, there is no correlation between X and Y.

Charles Edward Spearman (1863 - 1945)

Spearman was an English psychologist known for his work in statistics as a pioneer in factor analysis and for Spearman's rank correlation coefficient.

After serving for fifteen years as an officer in the British Army he went to Leipzig, Germany to study experimental psychology and obtained his Ph.D. degree in 1906. He was elected to Royal Society of London in 1924. In 1928 he became Professor of Psychology at University College, London. His many published papers cover a wide field, but he especially did pioneering work in the application of mathematical methods to the analysis of human mind. He discovered the general factor in human intelligence and developed a theory of 'g'.

He was greatly influenced by the work of Galton. He did pioneering work in psychology and developed correlation coefficient known by his name.

(a) Interpretation of $R = +1$ and $R = -1$

Two values of R need special attention. They are + 1 and – 1. R = + 1, when the scatter diagram is a straight line rising to the right. In this case the ranks of the value of X are the same as the ranks of the values of Y. $R = -1$, when the scatter diagram is a straight line falling to right. In this case, when the ranks of the values of X go on increasing in order, the ranks of the corresponding values of Y go on decreasing in the same order.

For example, consider the following data.

$R = +1$			
X	Y	R_1	R_2
8	115	1	1
11	120	2	2
14	125	3	3
17	130	4	4
20	135	5	5

$R = -1$			
X	Y	R_1	R_2
8	135	1	5
11	130	2	4
14	125	3	3
17	120	4	2
20	115	5	1

Note ....

When the scatter diagram is a straight line $r = +1$ or -1 and also $R = +1$ or -1.

(b) Relation between Spearman's Rank Correlation Coefficient R and Karl Pearson's Correlation Coefficient r.

Generally, for a given distribution the values of Spearman's rank correlation coefficient and Karl Pearson's correlation coefficient are different. Although both of them lie between + 1 and – 1, but actual values of the two coefficients for a given distribution are different. However, if the data are

such that if the values of the two variables x and y are arranged in either ascending or descending order and if they are found to increase or decrease by the equal amount *i.e.* if the difference between two values of x and the difference between the corresponding two values of y is constant, then the two values of R and r are equal. This is illustrated by the following example.

Example 1 : Calculate R and r from the following data.

X : 12 17 22 27 32.

Y : 113 119 117 115 121.

Interpret your result.

Sol. : **Calculation of R and r**

Sr. No.	$X - \bar{X}$			$Y - \bar{Y}$			xy	R_1	R_2	$(R_1 - R_2)^2$
	X	x	x^2	Y	y	y^2				
1	12	– 10	100	117	– 4	16	40	5	5	0
2	17	– 5	25	119	2	4	– 10	4	2	4
3	22	0	0	117	0	0	0	3	3	0
4	27	5	25	115	– 2	4	– 10	2	4	4
5	32	10	100	121	4	16	40	1	1	0
$N = 5$	110		250	585		40	60			8

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 8}{125 - 5} = 0{\cdot}6; \qquad \bar{X} = \frac{110}{5} = 22, \ \bar{Y} = \frac{585}{5} = 117$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}} = \frac{60}{\sqrt{250 \times 40}} = 0{\cdot}6.$$

Thus, the values of R and r are equal. It should be noted that the values of X increase by 5 and the values of Y, when arranged in ascending order also increase by the same amount 2 every time.

In general, if the values of x, when arranged in ascending order increase (or decrease) by a fixed amount and if the values of y, when arranged in ascending order increase (or decrease) by another (or the same) fixed amount, then the values of r and R come out to be equal.

(c) Computation of Correlation

There are two types of problems.

(i) When ranks of items are given.

(ii) When the actual values of the items are given.

(I) When ranks are given :

Steps : (i) Calculate the difference $D = R_1 - R_2$.

(ii) Calculate : D^2.

(iii) Apply the formula, $R = 1 - \frac{6\Sigma D^2}{N^3 - N}$.

(II) When the actual vales are given : We first ascertain the ranks of all items and follow the above procedure.

Example 1 : Compute Spearman's rank correlation coefficient from the following data.

X	:	18	20	34	52	12
Y	:	39	23	35	18	46

(M.U. 2016)

Sol. : First we give ranks to the data in descending order and then calculate $D^2 = (R_1 - R_2)^2$.

Sol. :

Calculation of R between X and Y

Serial No.	X	R_1	Y	R_2	D^2 $(R_1 - R_2)^2$
1	18	4	39	2	4
2	20	3	23	4	1
3	34	2	35	3	1
4	52	1	18	5	16
5	12	5	46	1	16
$N = 5$					$\Sigma D^2 = 38$

$$\therefore \quad R = 1 - \frac{6 \Sigma D^2}{N^3 - D}. \qquad \text{Here, } \Sigma D^2 = 38, \; N = 5.$$

$$\therefore \quad R = 1 - \frac{6 \times 38}{125 - 5} = 1 - \frac{228}{120} = 1 - 1{\cdot}9 = -0{\cdot}9$$

Example 2 : Calculate the rank correlation coefficient from the following data, relating to the ranks of 10 students in English and Mathematics.

Student No.	:	1,	2,	3,	4,	5,	6,	7,	8,	9,	10.
Rank in English	:	1,	3,	7,	5,	4,	6,	2,	10,	9,	8.
Rank in Mathematics	:	3,	1,	4,	5,	6,	9,	7,	8,	10,	2.

Sol. :

Calculation of R between English and Mathematics

Student No.	Rank in English R_1	Rank in Mathematics R_2	D^2 $(R_1 - R_2)^2$
1	1	3	4
2	3	1	4
3	7	4	9
4	5	5	0
5	4	6	4
6	6	9	9
7	2	7	25
8	10	8	4
9	9	10	1
10	8	2	36
$N = 10$			$\Sigma D^2 = 96$

Now, $R = 1 - \dfrac{6 \Sigma D^2}{N^3 - N}$ $\quad \because \; \Sigma D^2 = 96, \; N = 10$

$$\therefore \quad R = 1 - \frac{6 \times 96}{990} = 1 - \frac{96}{165} = 1 - 0{\cdot}5819 = 0{\cdot}4181$$

Example 3 : Calculate Spearman's coefficient of rank correlation from the data on height and weight of eight students.

Height (in inces)	:	60	62	64	66	68	70	72	74
Weight (in lbs.)	:	92	83	101	110	128	119	137	146

(M.U. 2016)

Sol. : **Calculation of R between Height and Weight**

Serial No.	Height	Rank R_1	Weight	Rank R_2	D^2 $(R_1 - R_2)^2$
1	60	1	92	2	1
2	62	2	83	1	1
3	64	3	101	3	0
4	66	4	110	4	0
5	68	5	128	6	1
6	70	6	119	5	1
7	72	7	137	7	0
8	74	8	146	8	0
$N = 8$					$\Sigma D^2 = 4$

Now, $R = 1 - \dfrac{6\Sigma D^2}{N^3 - N}$ $\quad \because \Sigma D^2 = 4,\ N = 8$

$\therefore \quad R = 1 - \dfrac{6 \times 4}{512 - 8} = 1 - \dfrac{24}{504} = 1 - 0{\cdot}048 = 0{\cdot}952$

(d) Equal Ranks

In some cases it may happen that there is a tie between two or more members *i.e.*, they have equal values and hence equal ranks. In such cases we divide the rank among equal members. For instance, if two items have 4th rank we divide the 4th and the next rank 5th between them equally and give $\dfrac{4+5}{2} = 4{\cdot}5$th rank to each of them. If three items have the same 4th rank, we give each of them $\dfrac{4+5+6}{3} = 5$th rank.

After assigning ranks in this way an adjustment is necessary. If m is the number of items having equal ranks then the factor $\dfrac{1}{12}(m^3 - m)$ is added to Σd_i^2. If there are more than one cases of this type this factor is added corresponding to each case. Then,

$$R = 1 - \frac{6\left[\Sigma d_i^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \ldots\right]}{n^3 - n}$$

Example 1 : Obtain the rank correlation coefficient from the following data.

X : 10, 12, 18, 18, 15, 40.

Y : 12, 18, 25, 25, 50, 25.

(M.U. 2004, 05, 10, 14, 19)

Sol. :

Calculation of R

X	Rank R_1	Y	Rank R_2	D^2 $(R_1 - R_2)^2$
10	1	12	1	0·00
12	2	18	2	0·00
18	4·5	25	4	0·25
18	4·5	25	4	0·25
15	3	50	6	9·00
40	6	25	4	4·00
$N = 6$				$\Sigma D^2 = 13·50$

There are two items in X series having equal values at the rank 4. Each is given the rank $= \frac{4+5}{2} = 4·5$. Similarly, there are three items in Y series at the rank 3. Each of them is given the rank $= \frac{3+4+5}{3} = 4$.

$$\therefore \quad R = 1 - \frac{6\left[\Sigma d_i^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2)\right]}{n^3 - n}$$

Since, $\Sigma D^2 = 13·50$, $m_1 = 2$, $m_2 = 3$, $N = 6$.

$$R = 1 - \frac{6\left[13·50 + \frac{1}{12}(8-2) + \frac{1}{12}(27-3)\right]}{216 - 6} = 1 - 0·4571 = 0·5429.$$

Example 2 : Calculate the value of rank correlation coefficient from the following data regarding marks of 6 students in statistics and accountancy in a test :

Marks in Statistics : 40, 42, 45, 35, 36, 39.

Marks in Accountancy : 46, 43, 44, 39, 40, 43. **(M.U. 2014, 19)**

Sol. :

Calculation of R

X	R_1	Y	R_2	D^2 $(R_1 - R_2)^2$
40	3	46	1	4·00
42	2	43	3·5	2·25
45	1	44	2	1·00
35	6	39	6	0·00
36	5	40	5	0·00
39	4	43	3·5	0·25
$N = 6$				$\Sigma D^2 = 7·50$

$$R = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(2^3 - 2)\right\}}{N^3 - N} = 1 - \frac{6(7·5 + 0·5)}{216 - 6}$$

$$= 1 - \frac{48}{210} = 1 - 0·229 = 0·771$$

Example 3 : From the following data calculate the coefficient of rank correlation between X and Y.

X : 32, 55, 49, 60, 43, 37, 43, 49, 10, 20.

Y : 40, 30, 70, 20, 30, 50, 72, 60, 45, 25.

Sol. : **Calculation of R between X and Y** **(M.U. 2016)**

X	R_1	Y	R_2	D^2 $(R_1 - R_2)^2$
32	3	40	5	4·00
55	9	30	3·5	30·25
49	7·5	70	9	2·25
60	10	20	1	81·00
43	5·5	30	3·5	4·00
37	4	50	7	9·00
43	5·5	72	10	20·25
49	7·5	60	8	0·25
10	1	45	6	25·00
20	2	25	2	0
$N = 10$				$\Sigma D^2 = 176$

Since there are two items in the X series having equal values at the rank 5 and two at the rank 7 they are given rank 5·5 and 7·5 each respectively. Similarly, in the Y series two items at the rank 3 are given the rank 3·5 each. There are three cases where there is a tie each having 2 times.

$$\therefore R = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \frac{1}{12}(m_3^3 - m_3)\right\}}{N^3 - N}$$

But, $\Sigma D^2 = 176$, $m_1 = m_2 = m_3 = 2$, $N = 10$

$$\therefore R = 1 - \frac{6\left\{176 + \frac{1}{12}(8-2) + \frac{1}{12}(8-2) + \frac{1}{12}(8-2)\right\}}{1000 - 10}$$

$$= 1 - \frac{6(177{\cdot}5)}{990} = 1 - 1{\cdot}076 = -0{\cdot}076.$$

EXERCISE - II

1. Sixteen industries of the State have been ranked as follows according to profits earned is 1980 - 81 and the working capital for the year. Calculate the rank correlation coefficient.

Industry :	A,	B,	C,	D,	E,	F,	G,	H,	I,	J,	K,	L,	M,	N,	O, P.
Rank (Profit) :	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11	12,	13,	14,	15, 16.
Rank (Capital) :	13,	16,	14,	15,	10,	12,	4,	11,	5,	9,	8,	3,	1,	6,	7, 2.

[**Ans.** : $R = -0{\cdot}8176$]

2. Distribution of marks in Economics and Mathematics for ten students in a certain test are given below :

Student No. :	1,	2,	3,	4,	5,	6,	7,	8,	9,	10.
Marks in Eco. :	25,	28,	32,	36,	38,	40,	39,	42,	41,	45.
Marks in Maths. :	70,	80,	85,	75,	59,	65,	48,	50,	54,	66.

Calculate the value of Spearman's Rank correlation coefficient.

[**Ans.** : – 0·6364]

3. Calculate Spearman's coefficient of rank correlation for the following data.

X : 53, 98, 95, 81, 75, 61, 59, 55.

Y : 47, 25, 32, 37, 30, 40, 39, 45. **[Ans. : – 0·9048]**

4. Calculate Spearman's coefficient of rank correlation from the following data.

X : 35, 38, 43, 30, 54, 68, 70, 92, 44, 56.

Y : 51, 37, 48, 62, 93, 73, 56, 72, 70, 92. **[Ans. : 0·59]**

5. Calculate Spearman's coefficient of rank correlation for the following data of scores in psychological tests (X) and arithmetical ability (Y) of 10 children.

Child : A, B, C, D, E, F, G, H, I, J.

X : 105, 104, 102, 101, 100, 99, 98, 96, 93, 92.

Y : 101, 103, 100, 98, 95, 96, 104, 92, 97, 94. **[Ans. : R = 0·782]**

6. Find the rank correlation coefficient between poverty and over crowding of cities from the following data.

Town	:	A, B, C, D, E, F, G, H, I, J.
No. of poor families	:	17, 13, 15, 16, 6, 11, 14, 9, 7, 12.
Population (over crowding)	:	30, 46, 35, 24, 12, 18, 27, 22, 46, 8.

[Ans. : R = 0·73]

7. Calculate the rank coefficient of correlation from the following data.

X : 105, 110, 112, 108, 111, 116, 120, 104, 115, 125.

Y : 39, 41, 45, 38, 48, 58, 60, 35, 54, 69. **[Ans. : R = 0·9636]**

8. Two judges X, Y ranked 8 candidates as follows. Find the correlation coefficient.

Candidates	:	A, B, C, D, E, F, G, H.
First judge X	:	5, 2, 8, 1, 4, 6, 3, 7.
Second judge Y	:	4, 5, 7, 3, 2, 8, 1, 6.

[Ans. : R = 0·67]

9. Calculate the rank correlation coefficient from the following data.

Marks in Paper I	:	52, 63, 45, 36, 72, 65, 45, 25.
Marks in Paper II	:	62, 53, 51, 25, 79, 43, 60, 33.

[Ans. : R = 0·648]

Miscellaneous Examples

Example 1 : State true or false with proper justification.

If coefficient of correlation between x and y is negative then the coefficient of correlation between $-x$ and $-y$ is positive. **(M.U. 1998)**

Sol. : Coefficient of correlation between x and y is given by

$$r = \frac{\Sigma xy - \Sigma x \cdot \Sigma y / N}{\sqrt{\Sigma x^2 - (\Sigma x)^2 / N}\sqrt{\Sigma y^2 - (\Sigma y)^2 / N}}$$

If we change the signs of x and y both, since the product and square terms occur, the sign of r will remain the same.

$\therefore$ The statement is false.

Example 2 : The coefficient of rank correlation of the marks obtained by 10 students in Physics and Chemistry was found to be 0·5. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation. **(M.U. 2005)**

Sol. : Since $R = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$ and $R = 0{\cdot}5,\ n = 10.$

$$0{\cdot}5 = 1 - \frac{6\Sigma d_i^2}{1000 - 10} \quad \therefore\ \Sigma d_i^2 = \frac{495}{6}$$

$\therefore$ Correct Σd_i^2 = Incorrect Σd_i^2 – (Incorrect rank diff.)^2 + (Correct rank diff.)^2

$$= \frac{495}{6} - 3^2 + 7^2 = \frac{735}{6}$$

$$\therefore \text{ Correct } R = 1 - \frac{6 \times (735/6)}{990} = 1 - \frac{735}{990} = 0{\cdot}26.$$

Example 3 : (a) Let $r_{xy} = 0{\cdot}4$, Cov. $(x, y) = 1{\cdot}6$, $\sigma_y^2 = 25$. Find σ_x.

(b) If $R_{x,y} = 0{\cdot}143$ and the sum of the squares of the differences between the ranks is 48, find N. **(M.U. 2005)**

Sol. : (a) We have $r = \frac{\text{cov}(x,y)}{\sigma_x \cdot \sigma_y}$. But $r = 0{\cdot}4$, cov. $(x, y) = 1{\cdot}6$, $\sigma_y = 5$.

$$\therefore\ 0{\cdot}4 = \frac{1{\cdot}6}{5\sigma_x} \quad \therefore\ \sigma_x = \frac{1{\cdot}6}{5 \times 0{\cdot}4} = 0{\cdot}8$$

(b) We have, $R = 1 - \frac{6\Sigma D^2}{N^3 - N}$. By data, $R = 0{\cdot}143,\ \Sigma D^2 = 48.$

$$\therefore\ 0{\cdot}143 = 1 - \frac{6 \times 48}{N^3 - N} = 1 - \frac{288}{N^3 - N} \quad \therefore\ \frac{288}{N^3 - N} = 1 - 0{\cdot}143 = 0{\cdot}857$$

$$\therefore\ N^3 - N = \frac{288}{0{\cdot}857} = 336 \quad \therefore\ N^3 - N - 336 = 0$$

$$\therefore\ N^3 - 7N^2 + 7N^2 - 49N + 48N - 336 = 0$$

$$\therefore\ (N-7)(N^2 + 7N + 48) = 0 \quad \therefore\ N = 7, \text{ other roots of } N \text{ are imaginary.}$$

Example 4 : Calculate the correlation coefficient between x and y from the following data.

$N = 10,\ \Sigma x = 140,\ \Sigma y = 150,\ \Sigma (x - 10)^2 = 180,$

$\Sigma (y - 15)^2 = 215,\ \Sigma (x - 10)(y - 15) = 60.$ **(M.U. 1997, 99)**

Sol. : With usual notation $\Sigma d_x^2 = 180,\ \Sigma d_y^2 = 215,\ \Sigma d_x d_y = 60.$

Now, $\bar{x} = A + \frac{\Sigma dx}{N} \quad \therefore\ 14 = 10 + \frac{\Sigma d_x}{10} \quad \therefore\ \Sigma d_x = 40$

Similarly, $\bar{y} = B + \frac{\Sigma d_y}{N} \quad \therefore\ 15 = 15 + \frac{\Sigma d_y}{10} \quad \therefore\ \Sigma d_y = 10$

Now, $$r = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \cdot \Sigma d_y}{N}}{\sqrt{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}\ \sqrt{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}}$$

$$\therefore \quad r = \frac{60 - \frac{40 \times 0}{10}}{\sqrt{180 - \frac{(40)^2}{10}}\sqrt{215 - \frac{(0)^2}{10}}} = \frac{60}{\sqrt{20}\sqrt{215}} = 0{\cdot}915.$$

EXERCISE - III

Type I

1. Compute Spearman's rank correlation coefficient from the following data.

   X : 85, 74, 85, 50, 65, 78, 74, 60, 74, 90

   Y : 78, 91, 78, 58, 60, 72, 80, 55, 68, 70. **(M.U. 2007, 18) [Ans. : $R = 0{\cdot}45$]**

2. Compute Spearman's rank correlation coefficient from the following data.

   X : 18, 20, 34, 52, 12.

   Y : 39, 23, 35, 18, 46. **(M.U. 2004) [Ans. : – 0·824]**

3. From the following data calculate Spearman's rank correlation between x and y.

   x : 36, 56, 20, 42, 33, 44, 50, 15, 60.

   y : 50, 35, 70, 58, 75, 60, 45, 80, 38. **(M.U. 2010) [Ans. : $R = 0{\cdot}92$]**

4. Find the coefficient of correlation (r) between x and y for the following data.

   x : 62 64 65 69 70 71 72 74

   y : 126 125 139 145 165 152 180 208 **(M.U. 2003, 04) [Ans. : 0·9032]**

5. Find Karl Pearson's coefficient of correlation and also, the Spear-man's rank coefficient of correlation for the following data.

   x : 12 17 22 27 32

   y : 113 119 117 115 121

   Also interpret your result. **(M.U. 2019) [Ans. : $r = R = 0{\cdot}6$]**

6. The following data gave the growth of employment in lakhs in organised sector in India between 1988 and 1995.

   Year : 1988, 89, 90, 91, 92, 93, 94, 95.

   Public Sector : 98, 101, 104, 107, 113, 120, 125, 128.

   Private Sector : 65, 65, 67, 68, 68, 69, 68, 68.

   Find the correlation coefficient (r) between the employment in public and private sectors and give your comments. **(M.U. 1998) [Ans. : $r = 0{\cdot}98$]**

7. Calculate the coefficient of correlation from the following figures. Is there any marked correlation between the production and price of tea ?

   Production in crores of lbs. : 44, 37, 31, 38, 36, 35, 40.

   Price in Rs. per lbs. : 2·75, 3·62, 4·25, 4·12, 4·28, 4·32, 4·05.

8. Draw a scatter diagram to represent the following data.

   X : 2, 4, 5, 6, 8, 11.

   Y : 18, 12, 10, 8, 7, 5.

   Calculate the coefficient of correlation between X and Y for the above data. **(M.U. 1998)**

   [Ans. : $r = -0{\cdot}92$]

9. Find the coefficient of correlation between height of father and height of son from the following data.

Height of father : 65, 66, 67, 67, 68, 69, 71, 73.

Height of son : 67, 68, 64, 68, 72, 70, 69, 70.

[**Ans.** : $r = +0{\cdot}55$]

10. Calculate Spearman's coefficient of rank correlation and Pearson's coefficient of correlation from the data on height and weight of eight students. Why the two values are same ?

Height ***(in inches)*** : 60, 62, 64, 66, 68, 70, 72, 74.

Weight ***(in lbs.)*** : 92, 83, 101, 110, 128, 119, 137, 146.

[**Ans.** : $r = R = 0{\cdot}93$; For both the series, the difference between consecutive terms remains constant if arranged in order.]

11. The following table shows the marks obtained by 10 students in Accountancy and Statistics. Find the Spearman's coefficient of rank correlation.

Student No. : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Accountancy : 45, 70, 65, 30, 90, 40, 50, 57, 85, 60.

Statistics : 35, 90, 70, 40, 95, 40, 60, 80, 80, 50.

Will the result change if the marks in the two subjects of all the students are increased by 5 and 10 respectively ? Will the result change if marks in the two subjects of all the students are halved ?

[**Ans.** : $r = 0{\cdot}8606$; No. : No.]

Type II

1. The coefficient of rank correlation between marks in Physics and Chemistry obtained by a group of students is 0·8. If the sum of the squares of differences in ranks in 33, find the number pairs students.

[**Ans.** : $N = 10$]

2. Find the number of pairs of observations from the following data.

$r = 0{\cdot}4$, $\Sigma xy = 108$, $\sigma_y = 3$, $\Sigma x^2 = 900$.

where x, y are the deviations of x, y from there respective means.

[**Ans.** : $N = 10$]

3. Coefficient of correlation between two variables is 0·4. Their covariance is 12. The variance of x is 25. Find the standard deviation of y.

[**Ans.** : $\sigma_y = 6$]

4. A computer while calculating the correlation coefficient between two variables x and y, from 25 observations obtained the following results

$N = 25$, $\Sigma x = 125$, $\Sigma y = 100$, $\Sigma x^2 = 650$, $\Sigma y^2 = 960$, $\Sigma xy = 508$

where x, y denote the actual values of the variables. Find the value of r.

[**Ans.** : 0·067]

5. A sample of 25 pairs of values of x and y lead to the following results.

$\Sigma x = 127$, $\Sigma y = 100$, $\Sigma x^2 = 760$, $\Sigma y^2 = 449$, $\Sigma xy = 500$.

Later on it was found that two pairs of values were taken as (8, 14) and (8, 6) instead of correct values (8, 12) and (6, 8).

Find corrected correlation coefficient between x and y.

(M.U. 2004) [**Ans.** : $r = -0{\cdot}31$]

6. Given : Number of pairs of observations = 10

X series standard deviation = 22·70. Y series standard deviation = 9·592

Summation of the products of corresponding deviations of X and Y from their respective actual means = – 1439. Find r.

[**Ans.** : $r = -0{\cdot}66$]

9. Regression and Fitting of Curves : Introduction

We have seen in the previous chapter how to examine and measure in magnitude and direction correlation between two variables. After establishing correlation, it is natural to search for a method which will help us to estimate the value of one variable when that of the other is known. This is achieved by the analysis of regression. Regression can be defined as **'a method of estimating the value of one variable when that of the other is known and when the variables are correlated'.**

The term, regression was first used by Galton. He found that although tall fathers have tall sons, and short fathers have short sons, the average height of sons of tall fathers is less than the average height of their fathers and the average height of sons of short fathers is more than the average height of their fathers. In other words the average height of sons of tall fathers or short fathers will regress or go back to the general average height. This phenomenon was described by him as 'regression.'

10. Lines of Regression

We have seen in the previous chapter that if the variables which are highly correlated are plotted on a graph then the points lie in a narrow strip. If the strip is nearly straight, we may draw a line such that all the points are close to it from both the sides. Such a line can be taken as the representative of the ideal variation. It is called the line of best fit. It is a line such that the sum of the distances of the points from the line is minimum. It is also called **'the line of regression'**. But we do not measure the distance by dropping a perpendicular from a point to the line. We measure, the deviations (i) vertically and (ii) horizontally, and get one line when distances are minimised vertically and second line when distances are minimised horizontally. Thus, we get **two lines of regression.**

(i) Line of regression of Y on X

If we minimise the deviations of the points from the line measured along y-axis we get a line which is called the **line of regression of Y on X**. Its equation is written in the from $Y = a + bX$. **This line is used for estimating the value of Y for a given value of X.** [See Fig. 2.1]

The equation of the line of regression of y on x must be written with y on the left hand side and x and the constant term on the right hand side.

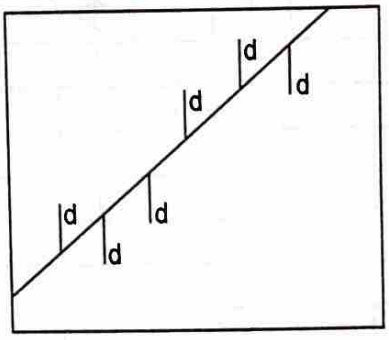


Fig. 2.1

For the Line of Regression of Y on X the distances d are minimised.

(ii) Line of regression of X on Y

If we minimise the deviations of the points from the line measured along x-axis we get a line which is called **the line of regression of X on Y.** Its equation is written in the form $X = a + bY$. **This line is used for estimating the value of X for a given value of Y.** [See Fig. 2.2]

The equation of the line of regression of x on y must be written with x on the left hand side and y and constant term on the right hand side.

There are two methods of obtaining the lines of regression. The first is graphical, the other is mathematical. They are :

1. The method of Scatter Diagram,
2. The Method of Least Squares.

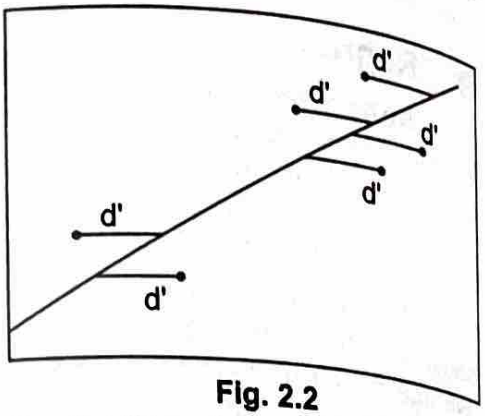


Fig. 2.2

For the line of regression of X on Y the distances d' are minimised.

11. The Method of Scatter Diagram

It is the simplest method of obtaining the lines of regression. The data are plotted on a graph paper by taking the independent variable on x-axis and the dependent variable on y-axis. We thus get points which are generally scattered. If the correlation is perfect *i.e.* if r is equal to one, positive or negative, the points will lie on a line, which is the line of regression. And there is only one line of regression and not two in such cases. However, in practice we rarely come across problems wherein we have perfect correlation. Usually, the points are scattered in a narrow straight strip and we have to find a line which will best represent all the points of the scatter diagram. We draw a line which will be close to all the points as far as possible.

Example : Given the following pairs of values of X and Y.

X : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12.

Y : 5, 6, 5, 6, 6, 8, 7, 9, 8, 9, 10, 11.

Plot the points on a graph and draw a line of regression.

Sol. :

Scatter Diagram and Line of Regression

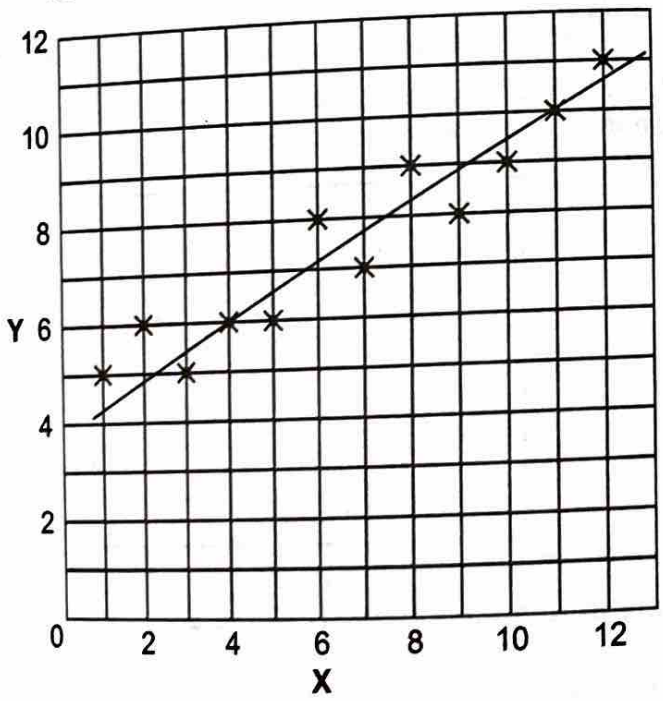


Fig. 2.3

12. The Method of Least Square

This is a mathematical method which gives an objective treatment to find a line of regression.

Let $y = a + bx$ be the equation of the line required. To find the line of regression of y on x we minimise the sum of the absolute distances of the points like $P(x_i, y_i)$ from the line measured along the y-axis. If Q is the point on the line corresponding to $P(x_i, y_i)$ we have to minimise the absolute distance PQ. Since Q lies on $y = a + bx$ its y-coordinate $= a + bx_i$

$$\therefore \quad |PQ| = |y_i - a - bx_i|$$

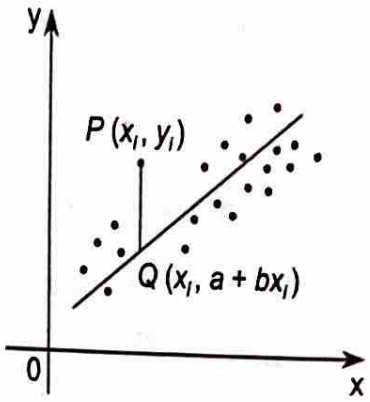


Fig. 2.4

For minimising $|PQ|$ we minimise its squares. Hence, if S denotes the sum of the squares of these distances,

$$S = \Sigma f_i (y_i - a - bx_i)^2 \qquad \text{where } f_i \text{ is the frequency of } (x_i, y_i).$$

We have to find a and b such that S is minimum, the conditions for which are

$$\frac{\partial S}{\partial a} = 2\Sigma f_i (y_i - a - b x_i) = 0 \quad \text{and} \quad \frac{\partial S}{\partial b} = 2\Sigma f_i (y_i - a - b x_i) x_i = 0$$

$$\Sigma f_i (y_i - a - bx_i) = 0 \qquad \ldots\ldots\ldots\ldots \text{(A)}$$

$$\Sigma x_i f_i (y_i - a - bx_i) = 0 \qquad \ldots\ldots\ldots\ldots \text{(B)}$$

From (A) we get, $\Sigma f_i y_i - a\Sigma f_i - b\Sigma f_i x_i = 0$

$$\therefore \quad N\bar{y} - aN - bN\bar{x} = 0 \quad \therefore \quad \bar{y} = a + b\bar{x} \qquad \ldots\ldots\ldots\ldots \text{(C)}$$

which shows that **the line of regression passes through** $(\bar{x}, \bar{y})$.

From (B) we get, $\Sigma f_i x_i y_i - a\Sigma f_i x_i - b\Sigma f_i x_i^2 = 0$ $\qquad \ldots\ldots\ldots\ldots \text{(D)}$

We now find the values of these expressions in terms of r, σ_x, σ_y.

But since, $r = \dfrac{1}{N} \dfrac{\Sigma f_i (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}$

$$r = \frac{\Sigma f_i x_i y_i - N\bar{x}\bar{y}}{N\sigma_x \sigma_y} \quad \therefore \quad \Sigma f_i x_i y_i = Nr\sigma_x\sigma_y + N\bar{x}\bar{y}$$

and $\sigma_x^2 = \dfrac{1}{N}\Sigma f_i (x_i - \bar{x})^2 = \dfrac{1}{N}\Sigma f_i x_i^2 - \bar{x}^2 \quad \therefore \quad \Sigma f_i x_i^2 = N\sigma_x^2 + N\bar{x}^2$

Putting the values of $\Sigma f_i x_i y_i$ and $\Sigma f_i x_i^2$ in D, we get

$$Nr\sigma_x\sigma_y + N\bar{x}\bar{y} = aN\bar{x} + bN\sigma_x^2 + bN\bar{x}^2$$

i.e. $\quad r\sigma_x\sigma_y + \bar{x}\bar{y} = a\bar{x} + b\sigma_x^2 + b\bar{x}^2 \qquad \ldots\ldots\ldots\ldots \text{(E)}$

Multiply (C) by $\bar{x}$ and subtract it form (E)

$$r\sigma_x\sigma_y = b\sigma_x^2 \quad \therefore \quad b = r\frac{\sigma_y}{\sigma_x}$$

Since, the line passes through $(\bar{x}, \bar{y})$ and its slope $b = r\dfrac{\sigma_y}{\sigma_x}$ its equation is

$$\boxed{y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})} \qquad \ldots\ldots\ldots\ldots \text{(F)}$$

Similarly, the equation of the line of regression of x on y can be shown to be

$$\boxed{x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y})} \quad \ldots\ldots\ldots\ldots (6)$$

Alternative Method : Instead of calculating $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ and r we may use the following method.

(i) The line of regression of y on x

Let the equation of the line of regression of y on x be $y = a + bx$.

Then as before we have to minimise

$S = \Sigma f_i(y_i - a - b x_i)^2$ the conditions of which are

$\frac{\partial S}{\partial a} = -2\Sigma f_i(y_i - a - bx_i) = 0$ and $\frac{\partial S}{\partial b} = -2\Sigma f_i(y_i - a - bx_i)\, x_i = 0$

i.e. $\Sigma f_i(y_i - a - bx_i) = 0$ and $\Sigma f_i x_i (y_i - a - bx_i) = 0$

i.e. $\Sigma f_i y_i - a\Sigma f_i - b\Sigma f_i x_i = 0$ and $\Sigma f_i x_i y_i - a\Sigma f_i x_i - b\Sigma f_i x_i^2 = 0$.

If the required sums are known, by solving the above two equations simultaneously for a and b we get the required equation.

In particular if all values occur only once *i.e.* if $f_i = 1$ for all i then the above equations take the form

$$\boxed{\Sigma y = aN + b\Sigma x} \quad \text{and} \quad \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2}$$

from which a and b can be calculated.

The above equations are called **normal equations**.

(ii) The line of regression of x on y

Let the equation of the line be $x = a + by$.

Proceeding as above we get the normal equations as

$$\boxed{\Sigma x = aN + b\Sigma y} \quad \text{and} \quad \boxed{\Sigma xy = a\Sigma y + b\Sigma y^2}$$

from which we can find the values of a and b.

Example : Find the equations of the lines of regression from the following data.

x:	5	6	7	8	9
y:	2	4	5	6	8

Also find r.

Sol. :

Calculations of regression

Sr. No.	x	x^2	y	y^2	xy
1	5	25	2	4	10
2	6	36	4	16	24
3	7	49	5	25	35
4	8	64	6	36	48
5	9	81	8	64	72
$N = 5$	35	255	25	145	189

The equation of the line of regression of y on x is $y = a + bx$ where a, b are given by

$$\Sigma y = aN + b\Sigma x \quad \text{and} \quad \Sigma xy = a\Sigma x + b\Sigma x^2$$

Putting the values of Σx, Σx^2, Σxy, we get

$25 = 5a + 35b$ ..... (i) and $189 = 35a + 255b$ ..... (ii)

Multiply the first by 7 and subtract it from the second.

$$\begin{array}{rl} 189 &= 35a + 255b \\ 175 &= 35a + 245b \\ \hline 14 &= \quad 10b \end{array} \qquad \therefore \ b = 1{\cdot}4$$

Putting this value of b in (i), we get

$25 = 5a + 35\,(1{\cdot}4)$ $\quad \therefore \ 5a = 25 - 49 = -24 \quad \therefore \ a = -4{\cdot}8$

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

$y = -4{\cdot}8 + 1{\cdot}4x$

The equation of the line of regression of x on y is $x = a + by$ where a, b are given by

$$\Sigma x = aN + b\Sigma y \quad \text{and} \quad \Sigma xy = a\Sigma y + b\Sigma y^2$$

Putting the values of Σx, Σy, Σxy, Σy^2, we get

$35 = 5a + 25b$ ..... (iii) and $189 = 25a + 145b$ ..... (iv)

Multiply (iii) by 5 and subtract it from (iv)

$$\begin{array}{rl} 189 &= 25a + 145b \\ 175 &= 25a + 120b \\ \hline 14 &= \quad 25b \end{array} \qquad \therefore \ b = 0{\cdot}56$$

Putting this value of b in (iii), we get

$35 = 5a + 25\,(0{\cdot}56)$ $\quad \therefore \ 5a = 35 - 14 = 11 \quad \therefore \ a = 2{\cdot}2$

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

$x = 2{\cdot}2 + 0{\cdot}56y$

Now, $r = \sqrt{b_1 \times b_2} = \sqrt{1{\cdot}4 \times 0{\cdot}56} = 0{\cdot}88$.

13. Calculations of the Equations of the Lines of Regression

There are various methods of calculating the equations of the lines of regression. The choice is yours. We state them below.

(a) By calculating the coefficient of correlation r and standard deviation σ_x and σ_y.

The equation of the line of regression of Y on X.

The equation of the line of regression of Y on X is given by,

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

where, $\overline{X}$ = the mean of X, $\overline{Y}$ = the mean of Y and $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$.

[See (F) and (G), page 2-25 and 2-26]

$\therefore$ The equation, therefore can be written as

$$\boxed{Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X})} \quad \ldots\ldots\ldots (1)$$

This equation is to be used for calculating the most probable value of Y for a given value of X.

The equation of the line of regression of X on Y.

The equation of the line of regression of X on Y is given by,

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

where, $\bar{X}$ = the mean of X, $\bar{Y}$ = the mean of Y and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$.

$\therefore$ The equation, therefore can be written as

$$\boxed{X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})} \quad \ldots\ldots\ldots (2)$$

This equation is to be used for calculating the most probable value of X for a given value of Y.

When the values of X and Y are known, we can as usual calculate r, σ_x and σ_y and then obtain equations of the lines of regression of X and Y.

(b) By calculating *a* and *b* directly

Instead of calculating $\bar{X}$, $\bar{Y}$, σ_x, σ_y and r we may calculate a and b directly as explained below.

The line of regression of Y on X : The constants a and b of the equation of the line of regression of Y on X *i.e.* of,

$$Y = a + bX$$

can be obtained by solving the following simultaneous equations.

$$\boxed{\Sigma Y = aN + \Sigma X} \quad \text{and} \quad \boxed{\Sigma YX = a\Sigma X + b\Sigma X^2} \quad \ldots\ldots\ldots (3)$$

The line of regression of X on Y : The constants a and b of the equation of the line of regression of X on Y *i.e.* of,

$$X = a + bY$$

can be obtained by solving the following simultaneous equations,

$$\boxed{\Sigma X = aN + b\Sigma Y} \quad \text{and} \quad \boxed{\Sigma XY = a\Sigma Y + b\Sigma y^2} \quad \ldots\ldots\ldots (4)$$

The equations, are called 'Normal equations'.

The formulae (3) and (4) are convenient when ΣX, ΣY, ΣXY, ΣX^2, ΣY^2 are known.

(c) By taking deviations from the means X and Y

If x and y denote the deviations of X and Y from their means, we know that,

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x\sigma_y} \cdot \frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{N\sigma_x^2} = \frac{\Sigma xy}{N \cdot \Sigma x^2/N} = \frac{\Sigma xy}{\Sigma x^2}$$

$\therefore$ $$\boxed{b_{yx} = \frac{\Sigma xy}{\Sigma x^2}} \quad \ldots\ldots\ldots (5)$$

Similarly, we can show that,

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} \quad \therefore \quad \boxed{b_{xy} = \frac{\Sigma xy}{\Sigma y^2}} \qquad \ldots\ldots\ldots\ldots (6)$$

where, $x = X - \bar{X}$, $y = Y - \bar{Y}$.

Hence, the equations of the lines of regression become,

$$\boxed{Y - \bar{Y} = \frac{\Sigma xy}{\Sigma x^2}(X - \bar{X})} \qquad \ldots\ldots\ldots\ldots (7)$$

and

$$\boxed{X - \bar{X} = \frac{\Sigma xy}{\Sigma y^2}(Y - \bar{Y})} \qquad \ldots\ldots\ldots\ldots (8)$$

(d) By taking deviations from assumed means

If the deviations of X and Y are taken from assumed means *i.e.* if $d_x = X - A$ and $d_y = Y - B$ then the coefficients b_{yx} and b_{xy} are given by,

$$\boxed{b_{yx} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{N}}{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}}} \qquad \ldots\ldots\ldots\ldots (9)$$

$$\boxed{b_{xy} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{N}}{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}}} \qquad \ldots\ldots\ldots\ldots (10)$$

(e) By using actual values directly

If X, Y are actual values of the two variates then it can be shown that

$$\boxed{b_{yx} = \frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}} \qquad \ldots\ldots\ldots\ldots (11)$$

$$\boxed{b_{xy} = \frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}} \qquad \ldots\ldots\ldots\ldots (12)$$

14. Regression Coefficients

The slope b of the line of regression of y on x *i.e.* b of the equation $y = a + bx$ is called the **coefficient of regression of y on x.** It represents the increment in y for unit change in the value of x. It is denoted by b_{yx}.

$\therefore$ b_{yx} = Coefficient of Regression of y on x.

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

Similarly, the slope b of the line of regression x on y *i.e.* b of the equation $x = a + by$ is called the **coefficient of regression of x on y**. It represents the increment in x for unit change in y. It is denoted by b_{xy}.

$\therefore$ b_{xy} = Coefficient of Regression of x on y.

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

Putting b_{yx} and b_{xy} which are the slopes of the lines of regression in (F) and (G), we can write the equations of lines of regression as

$$\boxed{y - \bar{y} = b_{yx}(x - \bar{x})} \quad \ldots\ldots\ldots (H)$$

and

$$\boxed{x - \bar{x} = b_{xy}(y - \bar{y})} \quad \ldots\ldots\ldots (I)$$

By putting the values of r and σ_y, σ_x in terms of actual values of x and y or by taking deviations from actual means or assumed means, we get the following formulae for b_{yx} and b_{xy}. (We repeat the first two.)

$$\boxed{b_{yx} = r\frac{\sigma_y}{\sigma_x}} \quad \ldots\ldots\ldots (1)$$

$$\boxed{b_{yx} = \frac{\Sigma xy}{\Sigma x^2}} \quad \ldots\ldots\ldots (2)$$

where $x = X - \bar{x}$, $y = Y - \bar{y}$ *i.e.* x, y are deviations of X, Y from actual means.

$$\boxed{b_{yx} = \frac{\Sigma d_x d_y - \dfrac{\Sigma d_x \cdot \Sigma d_y}{N}}{\Sigma d_x^2 - \dfrac{(\Sigma d_x)^2}{N}}} \quad \ldots\ldots\ldots (3)$$

where $d_x = X - A$, $d_y = Y - B$ *i.e.* d_x, d_y are deviations of X, Y from assumed means A and B.

$$\boxed{b_{yx} = \frac{\Sigma XY - \dfrac{\Sigma X \cdot \Sigma Y}{N}}{\Sigma X^2 - \dfrac{(\Sigma X)^2}{N}}} \quad \ldots\ldots\ldots (4)$$

where X, Y are the actual values of the variables.

Also,

$$\boxed{b_{xy} = r\frac{\sigma_x}{\sigma_y}} \quad \ldots\ldots\ldots (1')$$

$$\boxed{b_{xy} = \frac{\Sigma xy}{\Sigma y^2}} \quad \ldots\ldots\ldots (2')$$

where $x = X - \bar{x}$, $y = Y - \bar{y}$ *i.e.* x, y are the deviations of X, Y from actual means

$$\boxed{b_{xy} = \frac{\Sigma d_x d_y - \dfrac{\Sigma d_x \cdot \Sigma d_y}{N}}{\Sigma d_y^2 - \dfrac{(\Sigma d_y)^2}{N}}} \quad \ldots\ldots\ldots (3')$$

where $d_x = X - A$, $d_y = Y - B$ i.e. d_x, d_y are the deviations of X, Y from assumed means A and B.

$$b_{xy} = \frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}} \qquad \ldots\ldots\ldots\ldots (4')$$

where X, Y are the actual values of the variables.

15. Properties of Coefficients of Regression

1. Coefficient of correlation is the geometric mean between the coefficients of regression.

Proof : From the above results we have

$$b_{yx} \cdot b_{xy} = r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}$$

$$\therefore \quad \boxed{b_{yx} \cdot b_{xy} = r^2} \qquad \text{Hence, the result.}$$

Remark ....

Since the product of b_{yx} and b_{xy} is positive, if one of them is negative, the other also must be negative. In other words both the coefficients of regression are positive or both the coefficients of regression are negative together.

2. If one coefficient of regression is greater than one, the other must be less than one.

Proof : Since $-1 \le r \le 1$, $r^2 \le 1$.

Hence, from the above result,

$$b_{yx} \cdot b_{xy} \le 1 \quad \therefore \; b_{yx} \le \frac{1}{b_{xy}} \qquad \therefore \text{ If } b_{yx} < 1, \; b_{xy} > 1.$$

[See the values of b_{yx} and b_{xy} in Ex. 4, page 2-38. See that b_{yx} is less than one and b_{xy} is greater than one.]

3. Arithmetic mean of the coefficients of regression is greater than or equal to the coefficient of correlation.

Proof : We have to show that $\frac{b_{yx} + b_{xy}}{2} \ge r$

$$\text{i.e.} \quad \frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right) \ge r \quad \text{i.e.} \quad \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \ge 2$$

$$\text{i.e.} \quad \sigma_x^2 + \sigma_y^2 \ge 2\sigma_x\sigma_y \quad \text{i.e.} \quad \sigma_x^2 - 2\sigma_x\sigma_y + \sigma_y^2 \ge 0$$

$$\text{i.e.} \quad (\sigma_x - \sigma_y)^2 \ge 0 \text{ which is obviously true.}$$

Remark ....

In other words this means **the sum of the two coefficients of regression is greater than or equal to 2*r***. We shall verify this in Ex. 2, page 2-41; Ex. 5, page 2-39 and Ex. 6, page 2-43 below. We further note that equality will hold when $\sigma_x = \sigma_y$.

4. Coefficients of regression are independent of change of origin but not of change of scale.

If $u = ax + h$ and $v = by + k$, then

$$b_{uv} = \frac{a}{b} \cdot b_{xy} = \frac{\text{Coefficient of } x}{\text{Coefficient of } y} \cdot b_{xy}$$

We accept this result without proof.

5. If the correlation is perfect then the two coefficients of regression are reciprocals of each other.

Proof : We have $r = \pm 1$ and $r = \sqrt{b_{yx} \cdot b_{xy}}$ $\therefore$ $\pm 1 = \sqrt{b_{yx} \cdot b_{xy}}$.

Squaring $1 = b_{yx} \cdot b_{xy}$ $\therefore$ $b_{yx} = \frac{1}{b_{xy}}$.

e.g., if one coefficient of regression is 0·5 and if the correlation is perfect, then the other coefficient of regression is 2.

Example 1 : State whether the following statement is true or false with reasoning : "The regression coefficients between $2x$ and $2y$ are the same as those between x and y". **(M.U. 1997)**

Sol. : As seen above if $u = ax + h$ and $v = by + k$, $b_{uv} = \frac{a}{b} \cdot b_{xy}$.

But by data $u = 2x$ *i.e.* $a = 2$ and $v = 2y$ *i.e.* $b = 2$.

$\therefore$ $b_{vu} = \frac{2}{2} \cdot b_{xy} = b_{xy}$. Hence, the statement is true.

Example 2 : State whether the following statement is true or false : "The lines of regression between x and y are parallel to the lines of regression between $2x$ and $2y$.

Sol. : True. Explanation is left to you.

Example 3 : State whether the following statement is true or false : "The coefficients of regression between x and y are the same as the coefficients of regression between $2x + 5$ and $2y - 7$.

Sol. : True. Explanation is left to you.

Example 4 : If the arithmetic mean of regression coefficients is p and their difference is $2q$, find the correlation coefficient. **(M.U. 1998)**

Sol. : Let the coefficients of regression be b_1 and b_2.

Now by data $\frac{b_1 + b_2}{2} = p$ and $b_1 - b_2 = 2q$

$\therefore$ $b_1 + b_2 = 2p$ and $b_1 - b_2 = 2q$

$\therefore$ $b_1 = p + q$ and $b_2 = p - q$

$\therefore$ Coefficient of correlation $= r = \sqrt{b_1 b_2} = \sqrt{p^2 - q^2}$.

Example 5 : State true of false with reasoning : "$2x + y = 3$ and $x = 2y + 3$ cannot be the lines of regression." **(M.U. 2004)**

Sol. : If the first line is the line of regression of y on x it must written as $y = -2x + 3$ and if the second line is the line of regression of x on y, then it must be written as $x = 2y + 3$.

Hence, the coefficients of regression are $b_{yx} = -2$ and $b_{xy} = 2$ which is not possible as one of the them is negative and the other is positive and both are greater 1 numerically.

Now, we consider the lines in other way round. Let the first line be the line of regression of x on y and let the second line be the line of regression of y on x.

$$\therefore \quad x = -\frac{1}{2}y + \frac{3}{2} \quad \text{and} \quad y = \frac{1}{2}x - \frac{3}{2}.$$

Hence, the coefficients of regression are

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

which is again not possible because one is positive and the other is negative.

Hence, the statement is true.

Example 6 : State true or false with justification. If two lines of regression are $x + 3y - 5 = 0$ and $4x + 3y - 8 = 0$ then the correlation coefficient is + 0·5. **(M.U. 2003, 14)**

Sol. : Let the line $x + 3y - 5 = 0$ be the line of regression of x on y. Writing it as $x = -3y + 5$, we get

$$b_{xy} = -3.$$

Let the line $4x + 3y - 8 = 0$ be the line of regression of y on x. Writing it as $3y = -4x + 8$

i.e., as $y = -\frac{4}{3}x + 2$, we get $b_{yx} = -\frac{4}{3}$.

$$\therefore \quad r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{(-3)(-4/3)} = \sqrt{4} = 2$$

But r cannot be greater than 1.

Hence, our suppositions are wrong.

Now, let the line $x + 3y - 5 = 0$ be the line of regression of y on x. Writing it as

$$3y = -x + 5 \qquad i.e., \quad y = -\frac{1}{3}x + \frac{6}{3}, \text{ we get } \quad b_{yx} = -\frac{1}{3}.$$

Let the line $4x + 3y - 8 = 0$ be the line of regression of x on y. Writing it as

$$4x = -3y + 8 \quad i.e., \quad x = -\frac{3}{4}y + 2, \text{ we get } \quad b_{xy} = -\frac{3}{4}.$$

Now, $\quad r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\left(-\frac{1}{3}\right)\left(-\frac{3}{4}\right)} = \sqrt{\frac{1}{4}} = \frac{1}{2} = 0{\cdot}5$

Hence, the statement is true.

6. Angle between the lines of regression

The equation of the lines of regression of y on x is

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}). \qquad \text{Hence, its slope } m_1 = r\frac{\sigma_y}{\sigma_x}$$

The equation of the line of regression of x on y is

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

$$i.e., \quad y - \bar{y} = \frac{\sigma_y}{r \cdot \sigma_x}(x - \bar{x}) \qquad \text{Hence, its slope } m_2 = \frac{\sigma_y}{r \cdot \sigma_x}.$$

If θ is the angle between the lines of regression

$$\therefore \quad \tan\theta = \frac{m_1 - m_2}{1 + m_1 m_2} = \frac{r\frac{\sigma_y}{\sigma_x} - \frac{\sigma_y}{r\sigma_x}}{1 + \frac{r\sigma_y}{\sigma_x} \cdot \frac{\sigma_y}{r\sigma_x}} = \frac{\frac{(r^2\sigma_y\sigma_x - \sigma_y\sigma_x)}{r\sigma_x^2}}{\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}}$$

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

Corollary 1 : If $r = 0$, $\tan\theta = \infty \ \therefore \ \theta = \pi/2$.

The lines of regression are perpendicular to each other.

Corollary 2 : If $r = \pm 1$, $\tan\theta = 0 \ \therefore \ \theta = 0$.

The lines of regression are coincident.

Various Cases Are Shown Below.

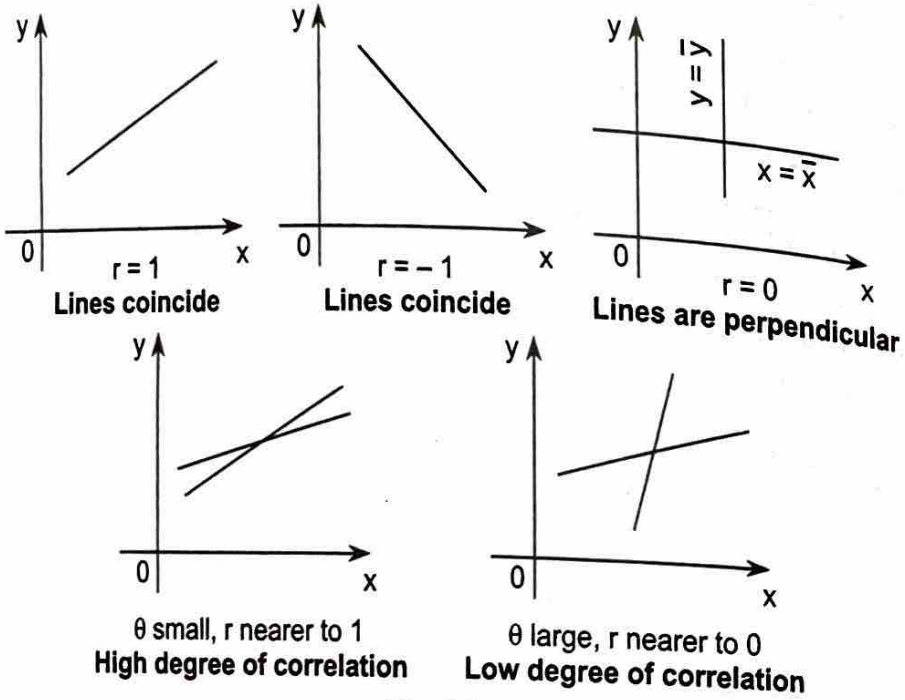


Fig. 2.5

Illustrative Examples

Example 1 : A panel of two judges A and B graded dramatic performances by independently awarding marks as follows :

Performance No.	:	1,	2,	3,	4,	5,	6,	7.
Marks by A	:	36,	32,	34,	31,	32,	32,	34.
Marks by B	:	35,	33,	31,	30,	34,	32,	36.

The eighth performance, however, which judge B could not attend, got 38 marks by judge A. If judge B had also been present, how many marks would he be expected to have awarded to the eighth performance ?

Sol. :

Calculations of coefficient of correlation etc.

Sr. No.	$X - \bar{X}$			$Y - \bar{Y}$			Product xy
	X	x	x^2	Y	y	y^2	
1	36	3	9	35	2	4	6
2	32	−1	1	33	0	0	0
3	34	1	1	31	−2	4	−2
4	31	−2	4	30	−3	9	6
5	32	−1	1	34	1	1	−1
6	32	−1	1	32	−1	1	1
7	34	1	1	36	3	9	3
$N = 7$	$\Sigma X = 231$		$\Sigma x^2 = 18$	$\Sigma Y = 231$		$\Sigma y^2 = 28$	$\Sigma xy = 13$

We have to find the marks that would have been awarded by the judge B. Therefore, let the marks given by the judge B be denoted by Y and those given by A by X.

$$\therefore \quad \bar{X} = \frac{\Sigma X}{N} = \frac{231}{7} = 33, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{231}{7} = 33$$

$$\therefore \quad b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{13}{18} = 0{\cdot}72$$

The equation of the line of regression of Y on X is,

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \quad i.e. \quad Y - 33 = 0{\cdot}72(X - 33)$$

To find the value of Y when $X = 38$, put $X = 38$ in the above equation.

$$\therefore \quad Y - 33 = 0{\cdot}72(38 - 43) = 0{\cdot}72 \times 5 = 3{\cdot}6$$

$$\therefore \quad Y = 33 + 3{\cdot}76 = 36{\cdot}6 = 37 \text{ approximately.}$$

$\therefore$ The judge B would have given 37 marks to the eighth performannce.

Alternatively :

Calculations of regression

Sr. No.	x	x^2	y	xy
1	36	1296	35	1260
2	32	1024	33	1056
3	34	1156	31	1054
4	31	961	30	930
5	32	1024	34	1088
6	32	1024	32	1024
7	34	1156	36	1224
$N = 7$	231	7641	231	7636

Let the marks given by A be x and those given by B be y.

Then the line of regression of y on x is $y = a + bx$.

And the normal equations are

$$\Sigma y = Na + b\Sigma x \qquad \therefore \quad 231 = 7a + 231b \qquad \ldots\ldots\ldots\ldots \text{(i)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \therefore \quad 7636 = 231a + 7641b \qquad \ldots\ldots\ldots\ldots \text{(ii)}$$

Multiply the first by 33 and subtract it from the second.

$$\therefore \quad 7636 = 231a + 7641b$$
$$7623 = 231a + 7626b$$
$$13 = 18b \qquad \therefore \quad b = 13 / 18 = 0{\cdot}72$$

Putting this value in (i), we get,

$$231 = 7a + 231\,(0{\cdot}72) \qquad \therefore \quad 7a = 231 - 231\,(0{\cdot}72) = 64{\cdot}68$$

$$\therefore \quad a = 9{\cdot}24$$

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

$$y = 9{\cdot}24 + 0{\cdot}72x$$

To estimate y when $x = 38$, we put $x = 38$ in the above equation

$$\therefore \quad y = 9{\cdot}24 + 0{\cdot}72\,(38) = 36{\cdot}6 = 37 \text{ approximately.}$$

Example 2 : Find the equations of the lines of regression for the following data.

x	:	5	6	7	8	9	10	11
y	:	11	14	14	15	12	17	16

(M.U. 2003, 05, 16, 17)

Sol. :

Calculations of regression

Sr. No.	x	x^2	y	y^2	xy
1	5	25	11	121	55
2	6	36	14	196	84
3	7	49	14	196	98
4	8	64	15	225	120
5	9	81	12	144	108
6	10	100	17	289	170
7	11	121	16	256	176
$N = 7$	56	476	99	1427	811

Now, the line of regression of y on x is $y = a + bx$.

The normal equations are

$$\Sigma y = Na + b \Sigma x \qquad \therefore \quad 99 = 7a + 56b \qquad \ldots\ldots\ldots (1)$$
$$\Sigma xy = a \Sigma x + b \Sigma x^2 \qquad \therefore \quad 811 = 56a + 476b \qquad \ldots\ldots\ldots (2)$$

Multiply the first equation by 56 and the second by 7 and subtract

$$\therefore \quad 5544 = 392a + 3136b$$
$$5677 = 392a + 3332b$$
$$-133 = -196b \qquad \therefore \quad b = \frac{133}{196} = 0{\cdot}6789$$

Putting this value of b in (1), we get

$$99 = 7a + 56 \times \frac{133}{196} \qquad \therefore \quad 7a = 99 - 38 \qquad \therefore \quad 7a = 61 \qquad \therefore \quad a = 8{\cdot}7143$$

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

$$y = 8{\cdot}7143 + 0{\cdot}6786x$$

Now, the equation of the line of regression of x on y is $x = a + by$.

The normal equations are

$\Sigma x = Na + b\Sigma y$ $\therefore$ $56 = 7a + 99b$ ..... (3)

$\Sigma xy = a\Sigma x + b\Sigma y^2$ $\therefore$ $811 = 99a + 1427b$ ..... (4)

Multiply the third equation by 99 and the fourth by 7 and subtract

$$\therefore \quad \begin{aligned} 5544 &= 693a + 9801b \\ 5677 &= 693a + 9989b \\ \hline 133 &= \qquad 188b \end{aligned} \qquad \therefore \quad b = \frac{133}{188} = 0{\cdot}7074$$

Putting this value of b in (1), we get

$$56 = 7a + 99 \times \frac{133}{188} \quad \therefore \quad 7a = 56 - 70{\cdot}0372 = -14{\cdot}0372 \quad \therefore \quad a = -2{\cdot}0053$$

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

$$x = -2{\cdot}0053 + 0{\cdot}7074y$$

(Further the coefficient of correlation is given by

$$r = \sqrt{b_1 b_2} = \sqrt{0{\cdot}6786 \times 0{\cdot}7074} = 0{\cdot}6928.)$$

Example 3 : From the following table showing age of cars of a certain make and annual maintenance costs, obtain the regression equation for costs related to age.

Age of Cars (years) : 2, 4, 6, 7, 8, 10, 12.

Annual maintenance Cost (Rs.) : 1,600, 1,500, 1,800, 1,900, 1,700, 2,100, 2,000.

Find the approximate cost of maintaining a 3 years old car of the same make.

Sol. :

Calculations of regression

Sr. No.	$X - \bar{X}$			$Y - \bar{Y}$			Product
	X	x	x^2	Y	y	y^2	xy
1	2	−5	25	1,600	−200	40,000	1000
2	4	−3	9	1,500	−300	90,000	900
3	6	−1	1	1,800	0	0	0
4	7	0	0	1,900	100	10,000	0
5	8	1	1	1,700	−100	10,000	−100
6	10	3	9	2,100	300	90,000	900
7	12	5	25	2,000	200	40,000	1000
$N = 7$	$\Sigma X = 49$		$\Sigma x^2 = 70$	$\Sigma Y =$ 12,600		$\Sigma y^2 =$ 2,80,000	$\Sigma xy = 3{,}700$

Let X denote age in years, Y denote cost in ₹

Now $\bar{X} = \frac{\Sigma X}{N} = \frac{49}{7} = 7$, $\bar{Y} = \frac{12600}{7} = 1{,}800$

and $b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{3700}{40} = 52{\cdot}86$

The equation of the line of regression of Y on X is,

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \quad \therefore \quad Y - 1800 = 52{\cdot}86\,(X - 7)$$

To find the value of Y when $X = 3$ put this value in the above equation,

$\therefore \quad Y - 1800 = 52{\cdot}86\,(3 - 7) = -211{\cdot}44$

$\therefore \quad Y = 1800 - 212.64 = 1588{\cdot}56$

$\therefore$ The cost of maintenance of 3 years old car = ₹ 1588·56.

Example 4 : Find the coefficients of regression and hence the equations of the lines of regression for the following data.

X : 78, 36, 98, 25, 75, 82, 90, 62, 65, 39.

Y : 84, 51, 91, 60, 68, 62, 86, 58, 53, 47.

Draw the lines of regression from your equations on the graph. Estimate the value of Y when $X = 50$ and the value of X when $Y = 90$ from the graph.

What is the significance of the point of intersection of the two lines?

Sol. : **Calculations of coefficients of regression**

Sr. No.	$X - \bar{X}$			$Y - \bar{Y}$			Product xy
	X	x	x^2	Y	y	y^2	
1	78	+ 13	169	84	+ 18	324	234
2	36	− 29	841	51	− 15	225	435
3	98	+ 33	1089	91	+ 25	625	825
4	25	− 40	1600	60	− 6	36	240
5	75	+ 10	100	68	+ 2	4	20
6	82	+ 17	289	62	− 4	16	− 68
7	90	+ 25	625	86	+ 20	400	500
8	62	− 3	9	58	− 8	64	24
9	65	0	0	53	− 13	169	0
10	39	− 26	676	47	− 19	361	494
$N = 10$	$\Sigma X = 650$		$\Sigma x^2 = 5398$	$\Sigma Y = 660$		$\Sigma y^2 = 2224$	$\Sigma xy = 2704$

(i) Now $\bar{X} = \dfrac{\Sigma X}{N} = \dfrac{650}{10} = 65, \quad \bar{Y} = \dfrac{\Sigma Y}{N} = \dfrac{660}{10} = 66.$

Coefficient of regression of Y on X is,

$$b_{yx} = \frac{\Sigma xy}{\Sigma x^2} = \frac{2704}{5398} = 0{\cdot}5009$$

Coefficient of regression of X on Y is,

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} = \frac{2704}{2224} = 1{\cdot}215$$

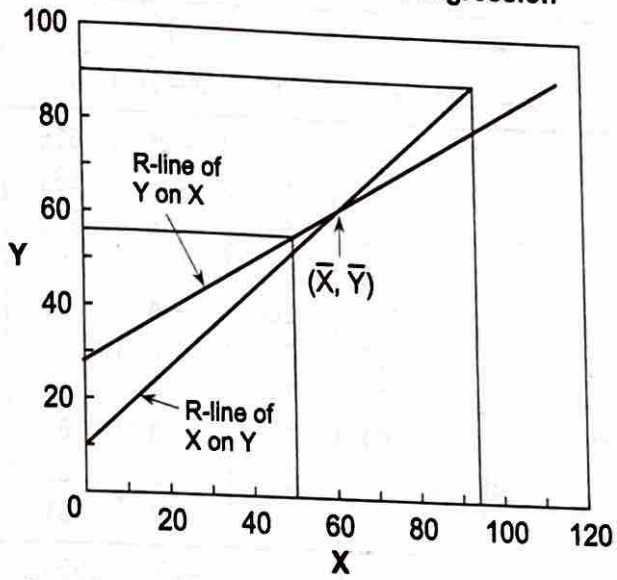
The equation of the line of regression of Y on X is,

$$Y - \bar{Y} = b_{yx}(X - \bar{X}) \qquad \therefore \quad Y - 66 = 0{\cdot}5\,(X - 65)$$

The equation of the line of regression of X on Y is,

$$X - \bar{X} = b_{xy}(Y - \bar{Y}) \qquad \therefore \quad X - 65 = 1{\cdot}2\,(Y - 66)$$

(II) **Graph of the two lines of regression**



To draw the lines of regression we take two points on each and connect them by a straight line.

From $Y - 66 = 0{\cdot}5\,(X - 65)$

we find that when $X = 65, \quad Y = 66$

and when $X = 0, \quad Y = 27{\cdot}5$

From $X - 65 = 1{\cdot}2\,(Y - 66)$

we find that when $X = 65, \quad Y = 66$

and when $X = 0, \quad Y = 12.$

To estimate Y when $X = 50$ we draw a line at $X = 50$, parallel to y-axis to meet the line of regression of Y on X and read the corresponding value of Y. It is 58. To estimate X when $Y = 90$, we draw a line at $Y = 90$, parallel to x-axis, to meet the line of regression of X on Y and read the corresponding value of X. It is 94.

The two lines intersect at the point for which $X = \bar{X} = 65$ and $Y = \bar{Y} = 66$.

Example 5 : A chemical engineer is investigating the effect of process operating temperature X on product yield Y. The study results in the following data.

X : 100, 110, 120, 130, 140, 150, 160, 170, 180, 190.

Y : 45, 51, 54, 61, 66, 70, 74, 78, 85, 89.

Find the equation of the least square line which will enable to predict yield on the basis of temperature. Find also the degree of relationship between the temperature and the yield.

(M.U. 2004, 16, 19)

Also verify that the sum of the coefficients of regression is greater than $2r$.

Sol. :

Calculations of b_{xy}, b_{yx} etc.

Sr. No.	dx			dy			$d_x d_y$
	X	$X-150$	d_x^2	Y	$X-70$	d_y^2	
1	100	−50	2500	45	−25	625	1250
2	110	−40	1600	51	−19	361	760
3	120	−30	900	54	−16	256	480
4	130	−20	400	61	−9	81	180
5	140	−10	100	66	−4	16	40
6	150	00	000	70	0	0	00
7	160	10	100	74	4	16	40
8	170	20	400	78	8	64	160
9	180	30	900	85	15	225	450
10	190	40	1600	89	19	361	760
$N=10$		−50	8500		−27	2005	4120

$$\bar{X} = A + \Sigma\frac{dx}{N} = 150 - \frac{50}{10} = 145; \qquad \bar{Y} = B + \Sigma\frac{dy}{N} = 70 - \frac{27}{10} = 67{\cdot}3$$

$$b_{yx} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{N}}{\Sigma d_x^2 - \frac{(\Sigma d_x)^2}{N}} = \frac{4120 - \frac{(-50)(-27)}{10}}{8500 - \frac{(-50)^2}{10}} = \frac{4120 - 135}{8500 - 250} = \frac{3985}{8250} = 0{\cdot}483$$

$$b_{xy} = \frac{\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{N}}{\Sigma d_y^2 - \frac{(\Sigma d_y)^2}{N}} = \frac{4120 - \frac{(-50)(-27)}{10}}{2005 - \frac{(-27)^2}{10}} = \frac{4120 - 135}{2005 - 72{\cdot}9} = \frac{3985}{1932{\cdot}1} = 2{\cdot}06$$

The line of regression of Y on X is

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

$\therefore \quad Y - 67{\cdot}3 = 0{\cdot}483\,(X - 145) \qquad \therefore \quad Y = 0{\cdot}483X - 2{\cdot}735$

The coefficient of correlation

$$r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0{\cdot}483 \times 2{\cdot}06} = 0{\cdot}9975$$

Now, $b_{xy} + b_{yx} = 2{\cdot}060 + 0{\cdot}483 = 2{\cdot}543$ and $2r = 2 \times 0{\cdot}9975 = 1{\cdot}995$

Hence, we see that $b_{yx} + b_{xy} > 2r$.

Miscellaneous Examples

Example 1 : Find the angle between the lines of regression using the following data.

$n = 10$, $\Sigma x = 270$, $\Sigma y = 630$, $\sigma_x = 4$, $\sigma_y = 5$, $r_{xy} = 0{\cdot}6$. **(M.U. 1998)**

Sol. : The angle between the lines of regression is given by

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

Putting the given values

$$\tan\theta = \left(\frac{1-0.6^2}{0.6}\right)\left(\frac{4\times 5}{16+25}\right) = 0.52.$$

Example 2 : Discuss the statement "The sum of the two coefficients of regression is always greater than $2r$ where r is the coefficient of correlation". **(M.U. 1998)**

Sol. : We have proved above (§ 15, 3, page 2-31) that

$$b_{yx} + b_{xy} \geq 2r.$$

Because this statement leads us to

$$(\sigma_x - \sigma_y)^2 \geq 0 \quad \text{which is always true.} \qquad \ldots\ldots\ldots\ldots\ldots \text{(A)}$$

This means the above given statement is partially true. The sum of the two coefficients of regression is greater than $2r$ but not always. The sum can be also equal to $2r$.

The condition (A) shows that if $\sigma_x = \sigma_y$, then the sign of equality will hold and then $b_{yx} + b_{xy}$ will be equal to $2r$.

This is also clear otherwise. From (1) and (1') (page 2-30), we have

$$b_{yx} + b_{xy} = r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}$$

If $\sigma_x = \sigma_y$, we will get $b_{yx} + b_{xy} = 2r$.

Example 3 : Given the following results of weights X and heights Y of 1000 men

$\overline{X} = 150$ lbs. $\quad \sigma_x = 20$ lbs.

$\overline{Y} = 68$ inches, $\quad \sigma_y = 2{\cdot}5$ inches, $\quad r = 0{\cdot}6$.

where $\overline{X}$ and $\overline{Y}$ are means of X and Y, σ_x and σ_y are standard deviations of X and Y and r is the correlation coefficient between X and Y.

John weighs 200 lbs., Smith is five feet tall. Estimate the height of John and weight of Smith. From the value of height of John estimate his weight. Why is it different from 200 ?

Sol. : With the given notation the line of regression of Y on X is

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

Substituting the given values,

$$Y - 68 = 0{\cdot}6 \times \frac{2{\cdot}5}{20}(X - 150) = \frac{15}{200}(X - 150).$$

Put $X = 200$,

$$\therefore \quad Y - 68 = \frac{15}{200}(200 - 150) = \frac{15}{4} = 3{\cdot}75$$

$$\therefore \quad Y = 68 + 3{\cdot}75 = 71{\cdot}75 \text{ inches.}$$

Now the line of regression of X of Y is

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

Substituting the given values,

$$X - 150 = 0{\cdot}6 \times \frac{20}{2{\cdot}5}(Y - 68) = \frac{24}{5}(Y - 68)$$

Put $Y = 5$ feet $= 60$ inches.

$$\therefore \quad X - 150 = \frac{24}{5}(60 - 68) = -\frac{192}{5}$$

$$\therefore \qquad X = 150 - \frac{192}{5} = 111{\cdot}6 \text{ lbs.}$$

Hence, height of John = 71·75 inches and weight of Smith = 111·6 lbs.

To estimate the weight of John from his height 71·25 we have to use the equation of line of regression of X on Y (and not Y on X).

i.e., $\quad X - 150 = \frac{24}{5}(Y - 68)$

Putting $Y = 71{\cdot}75$, we get

$$X - 150 = \frac{24}{5}(3{\cdot}75) = 18 \quad \therefore \; X = 168.$$

The difference is due to the fact that for estimating Y we use one equation and for estimating X we use another equation.

Example 4 : Given $6Y = 5X + 90$, $15X = 8Y + 130$, $\sigma_x^2 = 16$.

Find (i) $\bar{X}$ and $\bar{Y}$, (ii) r and (iii) σ_y^2. **(M.U. 2009, 10, 18, 19)**

Sol. : (i) To find $\bar{X}$ and $\bar{Y}$: We solve the given equations simultaneously. Multiply the first equation by 3.

$\therefore \; -15X + 18Y = 270$ and add $15X - 8Y = 130$

$\therefore \; 10Y = 40 \qquad \therefore \; \bar{Y} = 40$

Putting this value in any of the given equations.

$$6 \times 40 = 5X + 90 \quad \therefore \; X = 30 \quad \therefore \; \bar{X} = 30.$$

(II) **To find r :** Suppose the first equation represents, the line of regression of X on Y.

Writing it as $X = \frac{6}{5}Y - 18$, we find $b_{xy} = \frac{6}{5}$.

Suppose the second equation represents the line of regression of Y on X.

Writing it as $Y = \frac{15}{8}X - \frac{130}{8}$, we find $b_{yx} = \frac{15}{8}$.

$$\therefore \; r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{6}{5} \times \frac{15}{8}} = \sqrt{\frac{9}{4}} = \sqrt{2{\cdot}25} = 1{\cdot}5.$$

But the value of r can never be greater than 1 numerically. Hence, our supposition is wrong.

Now treating the first equation as representing the line of regression of Y on X, we write it as,

$$Y = \frac{5}{6}X + 15 \quad \therefore \; b_{yx} = \frac{5}{6}.$$

Treating the second equation as representing the line of regression of X on Y, we write it as,

$$X = \frac{8}{15}Y + \frac{130}{15} \quad \therefore \; b_{xy} = \frac{8}{15}$$

$$\therefore\ r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{8}{15} \times \frac{5}{6}} = \sqrt{\frac{4}{9}} = \frac{2}{3} = 0{\cdot}667.$$

(iii) To find σ_y. Consider, $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$ $\therefore\ \dfrac{5}{6} = \dfrac{2}{3} \times \dfrac{\sigma_y}{4}$ $\therefore\ \sigma_y = 5.$

Example 5 : The equations of the two regression lines are $3x + 2y = 26$ and $6x + y = 31$.
Find : (i) the means of x and y, (ii) coefficient of correlation between x and y, **(M.U. 2007, 16)**
(iii) σ_y if $\sigma_x = 3$.

Sol. : (i) To find $\bar{X}$ and $\bar{Y}$

We solve the equations simultaneously. Multiply the second by 2 and subtract from the first.

$$\therefore\quad \begin{array}{r} 3x + 2y = 26 \\ 12x + 2y = 62 \\ \hline 9x \qquad = 36 \end{array} \qquad \therefore\ x = 4.$$

Putting this value of x in the second equation, we get $24 + y = 31$ $\therefore\ y = 7.$

$\therefore\ \bar{X} = 4,\ \bar{Y} = 7.$

(ii) To find r : Suppose the first equation represents the line of regression of X on Y.

Writing it as $3x = -2y + 26.$ $\therefore\ x = -\dfrac{2}{3}y + \dfrac{26}{3}$

$\therefore$ We find that $b_{xy} = -\dfrac{2}{3}.$

Suppose the second equation represents the line of regression of Y on X.

Writing it as $y = -6x + 31$ $\therefore\ b_{yx} = -6.$

$$\therefore\ r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{(-2/3)(-6)} = \sqrt{4} = 2$$

But the value of r can never be greater than 1. Hence, our supposition is wrong.

Now, treating the first equation as representing the line of regression on of Y on X, we write it as

$$2y = -3x + 26 \qquad \therefore\ y = -\frac{3}{2}x + 13 \qquad \therefore\ b_{yx} = -\frac{3}{2}.$$

Treating the second equation as representing the line of regression of X on Y, we write it as

$$6x = -y + 31 \qquad \therefore\ x = -\frac{1}{6}y + \frac{31}{6} \qquad \therefore\ b_{xy} = -\frac{1}{6}.$$

$$\therefore\ r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\left(-\frac{3}{2}\right)\left(-\frac{1}{6}\right)} = \sqrt{\frac{1}{4}} = \frac{1}{2} = 0{\cdot}5$$

Since, both b_{xy} and b_{yx} are negative r is negative. $\therefore\ r = -0{\cdot}5.$

(iii) To find σ_y : Consider $b_{yx} = r\dfrac{\sigma_y}{\sigma_x}.$

But $b_{yx} = -\dfrac{3}{2}$, $r = -\dfrac{1}{2}$, and $\sigma_x = 3.$ $\therefore\ \sigma_y = b_{yx} \cdot \dfrac{\sigma_x}{r} = \left(-\dfrac{3}{2}\right) \cdot \dfrac{3}{(-1/2)} = 9$

Example 6 : The regression lines of a sample are $x + 6y = 6$, and $3x + 2y = 10$. Find (i) sample means $\bar{x}$ and $\bar{y}$, (ii) coefficient of correlation between x and y. Also estimate y when $x = 12$.
(M.U. 2004, 14, 15)

Also verify that the sum of the coefficients of regressions is greater than $2r$.

Sol. : (i) Mean $\bar{x}$ and $\bar{y}$ are obtained by solving the two given equations.

$$3x + 18y = 18 \quad \therefore \; y = 1/2$$
$$3x + 2y = 10 \quad \therefore \; x = 3$$
$$16y = 8$$

(ii) If the line $x + 6y = 6$ is the line of regression of y on x, then

$$6y = -x + 6 \;\; i.e. \;\; y = -\frac{1}{6}x + 1 \quad \therefore \; b_{yx} = -\frac{1}{6}$$

If the line $3x + 2y = 10$ is the line of regression of x on y, then

$$3x = -2y + 10 \;\; i.e. \;\; x = -\frac{2}{3}y + \frac{10}{3} \quad \therefore \; b_{xy} = -\frac{2}{3}$$

$$\therefore \; r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{\left(-\frac{1}{6}\right) \times \left(-\frac{2}{3}\right)} = \sqrt{\frac{1}{9}} = \frac{1}{3}$$

Since b_{yx} and b_{xy} are both negative, r is negative $\quad \therefore \; r = -1/3.$

Since $b_{yx} + b_{xy} = \frac{1}{6} + \frac{2}{3} = \frac{5}{6}$ (Numerically)

and $2r = \frac{2}{3}$, we see that $b_{yx} + b_{xy} > 2r$.

(iii) To estimate y when $x = 12$, we use the line of regression of y on x *i.e.* $y = -\frac{1}{6}x + 1$, when $x = 12$, $y = -2 + 1 = -1$.

Example 7 : If the tangent of the angle made by the line of regression of y on x is 0.6 and $\sigma_y = 2\,\sigma_x$, find the correlation coefficient between x and y. **(M.U. 2004, 09, 10, 15)**

Sol. : If the equation of the line of regression of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$ then we know that b_{yx} is the slope of the line of regression. We are thus, given $b_{yx} = 0{\cdot}6$.

But $\quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $\sigma_y = 2\sigma_x$

Putting these value, $\quad 0{\cdot}6 = r \cdot \frac{2\sigma_x}{\sigma_x} = 2r \quad \therefore \; r = \frac{0{\cdot}6}{2} = 0{\cdot}3.$

Example 8 : If $\sigma_x = \sigma_y = \sigma$ and the angle between the lines of regression is $\tan^{-1} 3$, find the coefficient of correlation. **(M.U. 2004)**

Sol. : We have

$$\tan\theta = \frac{1 - r^2}{r}\left(\frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}\right) \quad \therefore \; 3 = \frac{1 - r^2}{r} \cdot \left(\frac{\sigma^2}{\sigma^2 + \sigma^2}\right) = \frac{1 - r^2}{2r}$$

$$\therefore \; \frac{1 - r^2}{r} = 6 \quad \therefore \; r^2 + 6r - 1 = 0 \quad \therefore \; r = \frac{-6 \pm \sqrt{36 - 4}}{2} = -3 \pm 2\sqrt{2}$$

Since r cannot be numerically greater than 1, $r = -3 + 2\sqrt{2} = -0{\cdot}17$.

Example 9 : The following data regarding the heights (y) and weights (x) of 100 college students are given

$$\Sigma x = 15000, \; \Sigma x^2 = 2272500, \; \Sigma y = 6800, \; \Sigma y^2 = 463025, \; \Sigma xy = 1022250.$$

Find the coefficient of correlation between height and weight and also the equation of regression of height and weight. **(M.U. 1996)**

Sol. : The coefficients of regression are given by

$$b_{yx} = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{N}}{\Sigma x^2 - \frac{(\Sigma x)^2}{N}} = \frac{1022250 - \frac{15000 \times 6800}{100}}{2272500 - \frac{15000^2}{100}} = \frac{2250}{22500} = 0{\cdot}1.$$

$$b_{xy} = \frac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{N}}{\Sigma y^2 - \frac{(\Sigma y)^2}{N}} = \frac{1022250 - \frac{15000 \times 6800}{100}}{463025 - \frac{6800^2}{100}} = \frac{2250}{625} = 3{\cdot}6.$$

$\therefore \quad r = \sqrt{b_{yx} \times b_{xy}} = \sqrt{0{\cdot}1 \times 3{\cdot}6} = 0{\cdot}6.$

The equation of the lines of regression of y on x is

$y - \bar{y} = b_{yx}(x - \bar{x}) \quad \therefore \quad y - 68 = 0{\cdot}1(x - 1500) \quad \therefore \quad y = 0{\cdot}1x - 82.$

Example 10 : It is given that the means of x and y are 5 and 10. If the line of regression of y on x is parallel to the line $20y = 9x + 40$, estimate the value of y for $x = 30$. **(M.U. 1998, 2015)**

Sol. : The line of regression of y on x is $y - \bar{y} = b_{yx}(x - \bar{x})$.

Its slope is b_{yx}. But this line is parallel to $20y = 9x + 40$

i.e. $y = \frac{9}{20}x + 2$ whose slope is $\frac{9}{20}$. $\therefore \quad b_{yx} = \frac{9}{20}$.

But by data $\bar{x} = 5$ and $\bar{y} = 10$. Hence, the equation of the line of regression of y on x is

$$y - 10 = \frac{9}{20}(x - 5) \quad i.e. \quad y = \frac{9}{20}x + \frac{155}{20}$$

When $x = 30$, $y = \frac{270}{20} + \frac{155}{20} = \frac{425}{20} = 21{\cdot}25.$

Example 11 : If the arithmetic mean of regression coefficients is p and their difference is $2q$, find the correlation coefficient. **(M.U. 1998)**

Sol. : Let the coefficients of regression be b_1 and b_2.

Now by data $\frac{b_1 + b_2}{2} = p$ and $b_1 - b_2 = 2q$

$\therefore \quad b_1 + b_2 = 2p$ and $b_1 - b_2 = 2q$

$\therefore \quad b_1 = p + q$ and $b_2 = p - q$

$\therefore$ Coefficient of correlation $= r = \sqrt{b_1 b_2} = \sqrt{p^2 - q^2}$

EXERCISE - IV

Type I

State true or false with proper reasoning.

1. If $r = 0$, the lines of regression are parallel to each other. **[Ans. : False]**
2. The values of r and R can never be equal. **[Ans. : False]**

3. In a regression analysis it was found that $b_{yx} = 0{\cdot}87$, $b_{xy} = 1{\cdot}55$. These values are not consistent. [**Ans.** : True]

4. The two regression coefficients are both positive or both negative. [**Ans.** : True]

5. $3x + y = 5$ and $2x - 3y = 7$ cannot be lines of regression for any set of values of x and y. [**Ans.** : True]

Type II

1. The following table gives the age of car of a certain make and annual maintenance cost. Obtain the equation of the line of regression of cost on age.

Age of a car	:	2	4	6	8
Maintenance	:	1	2	2·5	3

(M.U. 1998, 2014) [**Ans.** : $y = 0{\cdot}325x + 0{\cdot}5$]

2. Obtain the equation of the line of regression of y on x from the following data and estimate y for $x = 73$.

x	:	70,	72,	74,	76,	78,	80.
y	:	163,	170,	179,	188,	196,	220.

[**Ans.** : $y = 5{\cdot}31x - 212.57$; $y = 175{\cdot}37$] **(M.U. 1997)**

3. The heights in cms of fathers (x) and of the eldest sons (y) are given below.

x	:	165	160	170	163	173	158	178	168	173	170	175	180
y	:	173	168	173	165	175	168	173	165	180	170	173	178

Estimate the height of the eldest son if the height of the father is 172 cms. and the height of the father if the height of the eldest son is 173 cm.

Also find the coefficient of correlation between the heights of fathers and sons.

[**Ans.** : **(i)** $y = 1{\cdot}016x - 5{\cdot}123$, **(ii)** $x = 0{\cdot}476y + 98{\cdot}98$, **(iii)** 169·97, 173·45, **(iv)** $r = 0{\cdot}696$] **(M.U. 2002, 05)**

4. Find (i) the lines of regression, (ii) coefficient of correlation for the following data.

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

[**Ans.** : $y = 19{\cdot}79 + 0{\cdot}72x$, **(ii)** $x = 33{\cdot}29 + 0.5y$; $r = 0{\cdot}6$] **(M.U. 2002, 14, 19)**

5. Find the line of regression for the following data and estimate y corresponding to $x = 15{\cdot}5$.

x	:	10	12	13	16	17	20	25
y	:	19	22	24	27	29	33	37

(M.U. 2004, 19) [**Ans.** : $y = 0{\cdot}8x + 13{\cdot}23$; 25·63]

Type III

1. Given

	x series	**y series**	
Mean	18	100	
S.D.	14	20	$r = 0{\cdot}8$.

Find the most probable value of y when $x = 70$ and most probable value of x when $y = 90$.

[**Ans.** : $y = 159{\cdot}3$, $x = 12{\cdot}4$]

2. Given the following information about marks of 60 students.

	Mathematics	**English**
Mean	80	50
S.D.	15	10

Coefficient of correlation $r = 0{\cdot}4$. Estimate the marks of the student in mathematics who scored 60 marks in English. **(M.U. 2006)** [**Ans. :** 86]

3. You are supplied with the following information. The equation of the lines regression are $2x + 3y + 8 = 0$ and $x + 2y - 5 = 0$.
Find the means of x and y and the coefficient of correlation between them. **(M.U. 1997)**
[**Ans. :** $\bar{x} = -31$, $\bar{y} = 18$, $r = -0{\cdot}87$]

4. From 8 observations the following results were obtained :
$\Sigma x = 59$, $\Sigma y = 40$, $\Sigma x^2 = 524$, $\Sigma y^2 = 256$, $\Sigma xy = 364$.
Find the equation of the line of regression of x on y and the coefficient of correlation.
(M.U. 2018) [**Ans. :** $x = 1{\cdot}5y - 0{\cdot}5$, $r = 0{\cdot}98$]

5. The equations of the two lines of regression are $x = 19{\cdot}13 - 0{\cdot}87y$ and $y = 11{\cdot}64 - 0{\cdot}50x$.
Find (i) the means of x and y, (ii) the coefficient of correlation between x and y.
(M.U. 2004) [**Ans. : (i)** $\bar{x} = 15{\cdot}79$, $\bar{y} = 3{\cdot}74$; **(ii)** $r = -0{\cdot}66$, $b_{yx} = -0{\cdot}50$, $b_{xy} = -0{\cdot}87$]

6. Out of the two equations given below which can be a line of regression of x on y and why ?
$x + 2y - 6 = 0$ and $2x + 3y - 8 = 0$. **(M.U. 2003)** [**Ans. :** $2x + 3y - 8 = 0$]

7. In a partially destroyed laboratory record of analysis of correlation data the following results are legible. Variance of $x = 9$, equations of the lines of regression
$4x - 5y + 33 = 0$, $20x - 9y - 107 = 0$.
Find (i) the mean values of x and y, (ii) the standard deviation of y, (iii) coefficient of correlation between x and y. **(M.U. 1999, 2003, 19)** [**Ans. : (i)** $\bar{x} = 13$, $\bar{y} = 17$, **(ii)** $\sigma_y = 4$, **(iii)** $r = 0{\cdot}6$]

8. Given : var $(x) = 25$. The equations of the two lines of regression are $5x - y = 22$ and $64x - 45y = 24$. **(M.U. 1998)**
Find (i) $\bar{x}$ and $\bar{y}$, (ii) r, (iii) σ_y.
[**Ans. : (i)** $\bar{x} = 6$, $\bar{y} = 8$, **(ii)** $r = 1{\cdot}87$, **(iii)** $\sigma_y = 1/5$]

9. Find the regression coefficients and the coefficient of correlation from the following data where x, y denote the actual values.
$N = 12$, $\Sigma x = 120$, $\Sigma y = 432$, $\Sigma xy = 4992$, $\Sigma x^2 = 1392$, $\Sigma y^2 = 18252$.
[**Ans. :** $b_{yx} = 3{\cdot}5$, $b_{xy} = 0{\cdot}249$, $r = 0{\cdot}93$]

16. Fitting of Curves : Introduction

In many social, economical, engineering and physical problems we have a set of values of x and y although we do not know the functional relationship between them. Fitting of a curve to a given set of values means finding a functional relationship between x and y whose curve is the closest possible curve to the given values. The curve so obtained does not pass through all the given points but is close to them to the maximum extent. Finding such a curve for a given set of values is called **Curve Fitting.** The relation in general is assumed to be a linear function $y = a + bx$ or a parabolic function $y = a + bx + cx^2$ or even exponential or logarithmic. The method used for fitting the curve is based on the **principle of least squares.**

17. Fitting a Straight Line by the Method of Least Squares

Suppose we have a set of values (x_i, y_i). Suppose further that we want to fit a straight line $y = a + bx$ to these values. The straight line must be close to the given points to the maximum extent. **The principle of least square states that the straight line should be such that the distances of the given points from the straight line measured along the y-axis must be minimum.** The line obtained in this way is called **the line of best fit.**

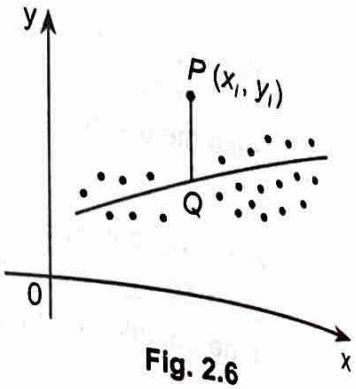


Fig. 2.6

Suppose $P(x_i, y_i)$ is a given point and suppose the equation of the line of best fit be $y = a + bx$. Suppose further that the line parallel to the y-axis through P intersects the line in Q. Now, the coordinates of Q are $(x_i, a + bx_i)$. We want to find a, b such that the distance $|PQ|$ is minimum. But the distance $|PQ|$ is minimum when the square of the distance $(y_i - a - bx_i)^2$ is minimum. This must be true for all points. This means we should have $S \equiv \Sigma (y_i - a - bx_i)^2$ minimum. Since we find a, b such that the sum of the squares S is least, the method is known as the method of least squares.

Now, for S to be minimum the conditions are

$$\frac{\partial S}{\partial a} = 0 \quad \text{and} \quad \frac{\partial s}{\partial b} = 0$$

$$\therefore \quad \Sigma (y_i - a - bx_i) = 0 \quad \text{and} \quad \Sigma (y_i - a - bx_i)\, x_i = 0 \qquad \therefore \quad \Sigma y_i = \Sigma a - b \Sigma x_i$$

$$\therefore \quad \Sigma y_i = Na + b \Sigma x_i \quad \text{and} \quad \Sigma y_i x_i = a \Sigma x_i + b \Sigma x_i^2$$

We shall drop the suffix i and write the equations as

$$\boxed{\begin{aligned} \Sigma y &= Na + b \Sigma x \\ \Sigma xy &= a \Sigma x + b \Sigma x^2 \end{aligned}}$$

These equations are known as **Normal Equations.**

Methods of finding the constants a

1. Direct Method : If we find Σx, Σy, etc. and substitute their values in the above equations, we get two equations in two unknowns a and b. Solving these two equations we can find the constants a and b.

2. Assumed Mean Method : We may assume a certain mean of x and of y, *i.e.*, we may put $X = x - m$ and $Y = y - n$ and find ΣX, ΣY, etc. put these values in the above two equations.

If necessary, we may divide the deviations by a common factor, if there is any, *i.e.*, we may put $X = \frac{x - m}{h}$ and $Y = \frac{y - n}{k}$.

Solving these equations, we get a and b and resubstituting the values of X and Y, we get the equation of the line.

3. Actual Mean Method : If actual means of x and y are round figures, we put $X = x - \bar{x}$ and $Y = y - \bar{y}$ $\left(\text{or } X = \frac{x - \bar{x}}{h},\ Y = \frac{y - \bar{y}}{k}\right)$, find ΣX, ΣY, etc. and put these values in the above two equations.

Solving these equations as above and resubstituting the values of X and Y, we get the equation of the line.

Of these three methods we shall see below that the last method is the simplest one.

Example 1 : Fit a straight line to the following data :

(x, y) : (1, 1), (2, 5), (3, 11), (4, 8), (5, 14)

Sol. : We shall obtain the line from the above normal equations.

Calculations of Σx, Σx^2 etc.

x	y	x^2	xy
1	1	1	1
2	5	4	10
3	11	9	33
4	8	16	32
5	14	25	70
$N = 5, 15$	39	55	146

The normal equations are

$$\Sigma y = Na + b \Sigma x \quad \therefore \; 39 = 5a + 15b \quad \ldots\ldots (1)$$

and $$\Sigma xy = a \Sigma x + b \Sigma x^2 \quad \therefore \; 146 = 15a + 55b \quad \ldots\ldots (2)$$

Multiply (1) by 3,

$\therefore \; 117 = 15a + 45b$ and $146 = 15a + 55b$

By subtraction $10b = 29$ $\therefore \; b = 2{\cdot}9.$

Putting this value of b in (1),

$$5a = 39 - 15 \times 2{\cdot}9 = -4{\cdot}5$$

$$\therefore \; a = -\frac{4{\cdot}5}{5} = -0{\cdot}9.$$

Hence, the equation of the line is $y = -0{\cdot}9 + 2{\cdot}9x.$

Example 2 : Fit a straight line $y = a + bx$ to the following data.

x	:	0	5	10	15	20	25
y	:	12	15	17	22	24	30

Sol. : (1) Direct Method

Calculations of Σx, Σx^2 etc.

x	y	x^2	xy
0	12	0	0
5	15	25	75
10	17	100	170
15	22	225	330
20	24	400	480
25	30	625	750
$N = 6, \Sigma x = 75$	$\Sigma y = 120$	$\Sigma x^2 = 1375$	$\Sigma xy = 1805$

Let the equation be $y = a + bx$.

The normal equations are

$\Sigma y = Na + b\Sigma x$ $\quad\therefore\quad 120 = 6a + 75b$ ..... (1)

and $\Sigma xy = a\Sigma x + b\Sigma x^2$ $\quad\therefore\quad 1805 = 75a + 1375b$ ..... (2)

Now, to find a, b multiply (1) by 50 and (2) by 4 and subtract.

$$\begin{array}{rcl} 6000 & = & 300\,a + 3750\,b \\ 7220 & = & 300\,a + 5500\,b \\ \hline 1220 & = & 1750\,b \end{array}$$

$$\therefore\quad b = \frac{1220}{1750} = 0{\cdot}6971 = 0{\cdot}7$$

Putting this value of b in (1), we get

$120 = 6a + 75\,(0{\cdot}7)$ $\quad\therefore\quad 6a = 120 - 52{\cdot}5 = 67{\cdot}5$ $\quad\therefore\quad a = 11{\cdot}25.$

$\therefore$ The equation of the line is $y = 11{\cdot}25 + 0{\cdot}7x$.

(2) Assumed Mean Method : We shall assume 10 and 17 as means of x and y, *i.e.*, we shall put $X = x - 10$ and $Y = y - 17$.

Calculations of ΣX, ΣX^2 etc.

x	y	$X = x - 10$	$Y = y - 17$	X^2	XY
0	12	– 10	– 5	100	50
5	15	– 5	– 2	25	10
10	17	0	0	0	0
15	22	5	5	25	25
20	24	10	7	100	70
25	30	15	13	225	195
$N = 6$		15	18	475	350

Let the equation be $Y = a + bX$.

The normal equations are

$\Sigma Y = Na + b\Sigma X$ $\quad\therefore\quad 18 = 6a + 15b$ ..... (1)

and $\Sigma XY = a\Sigma X + b\Sigma X^2$ $\quad\therefore\quad 350 = 15a + 475$ ..... (2)

Multiply (1) by 5 and (2) by 2 and subtract.

$$\therefore\quad \begin{array}{rcl} 90 & = & 30\,a + 75\,b \\ 700 & = & 30\,a + 950\,b \\ \hline 610 & = & 875\,b \end{array}$$

$$\therefore\quad b = \frac{610}{875} = 0{\cdot}7$$

Putting this value of b in (1), we get

$18 = 6a + 15\,(0{\cdot}7)$ $\quad\therefore\quad 6a = 18 - 10{\cdot}5$ $\quad\therefore\quad a = 1{\cdot}25.$

$\therefore$ The equation is $Y = 1{\cdot}25 + 0{\cdot}7X$.

Resubstituting $Y = y - 17$ and $X = x - 10$, we get

$y - 17 = 1{\cdot}25 + 0{\cdot}7\,(x - 10)$ $\quad\therefore\quad y = 0{\cdot}7x + 11{\cdot}25.$

Example 3 : Fit a straight line $y = a + bx$ to the following data.

x :	1	2	3	4	5	6
y :	49	54	60	73	80	86

(M.U. 2014)

Sol. : We shall solve this example by all the three methods.

(a) Direct method :

Calculations of Σx, Σx^2 etc.

x	y	x^2	xy
1	49	1	49
2	54	4	108
3	60	9	180
4	73	16	292
5	80	25	400
6	86	36	516
$N = 6$, 21	402	91	1545

Now, $y = a + bx$

$\Sigma y = Na + b\Sigma x$ $\therefore$ $402 = 6a + 21b$ ..... (1)

and $\Sigma xy = a\Sigma x + b\Sigma x^2$ $\therefore$ $1545 = 21a + 91b$ ..... (2)

Multiply (1) by 7 and (2) by 2 and subtract

$\therefore$ $2814 = 42a + 147b$

$3090 = 42a + 182b$

$276 = 35b$ $\therefore$ $b = 7{\cdot}89$

Putting this value of b in (1), we get

$402 = 6a + 21\ (7{\cdot}89)$

$\therefore$ $6a = 402 - 165{\cdot}69 = 236{\cdot}31$ $\therefore$ $a = 39{\cdot}38$

$\therefore$ $y = 39{\cdot}38 + 7{\cdot}89x$.

(b) Assumed Mean Method : We shall assume 3 and 60 as means of x and y and find the deviations.

Calculations of ΣX, ΣX^2 etc.

x	y	$X = x - 3$	$Y = y - 60$	X^2	XY
1	49	-2	-11	4	22
2	54	-1	-6	1	6
3	60	0	0	0	0
4	73	1	13	1	13
5	80	2	20	4	40
6	86	3	26	9	78
$N = 6$		3	42	19	159

Now, $Y = a + bX$

$\Sigma Y = Na + b\Sigma X$ $\therefore$ $42 = 6a + 3b$ ..... (1)

and $\Sigma XY = a\Sigma X + b\Sigma X^2$ $\therefore$ $159 = 3a + 19b$ ..... (2)

Multiply (2) by 2 and subtract (1) from the result.

$\therefore \quad 318 = 6a + 38b$

$42 = 6a + 3b$

$276 = 35b \qquad \therefore \quad b = 7{\cdot}89$

Putting this value of b in (1),

$42 = 6a + 3(7{\cdot}89) \qquad \therefore \quad 6a = 18{\cdot}33 \qquad \therefore \quad a = 3{\cdot}05$

$\therefore \quad Y = 3.05 + 7.89\,X.$

Putting $X = x - 3$, $Y = y - 60$, we get

$y - 60 = 3{\cdot}05 + 7{\cdot}89\,(x - 3)$

$\therefore \quad y = 39{\cdot}38 + 7{\cdot}89\,x.$

(c) Actual Mean method : We shall take deviations from actual means of x and y.

Now, $\bar{x} = \frac{21}{6} = 3{\cdot}5, \quad \bar{y} = \frac{402}{6} = 67.$

Calculations of ΣX, ΣX^2 etc.

x	y	$X = x - 3{\cdot}5$	$Y = y - 67$	X^2	XY
1	49	− 2·5	− 18	6·25	45·0
2	54	− 1·5	− 13	2·25	19·5
3	60	− 0·5	− 7	0·25	3·5
4	73	0·5	6	0·25	3·0
5	80	1·5	13	2·25	19·5
6	86	2·5	19	6·25	47·5
$N = 6$		0	0	17·5	138·0

Now, $Y = a + bX$

$\therefore \quad \Sigma Y = Na + b\Sigma X \qquad$ and $\qquad \Sigma XY = a\Sigma X + b\Sigma X^2$

Putting the above values

$0 = 6a + b\,(0) \qquad \therefore \quad a = 0$

and $\quad 138 = a\,(0) + 17{\cdot}5\,b \qquad \therefore \quad b = 7{\cdot}89$

Hence, the equation is $\quad Y = 7{\cdot}89X.$

Putting the values of X, Y, we get

$y - 67 = 7{\cdot}89\,(x - 3{\cdot}5) \qquad \therefore \quad y = 39{\cdot}38 + 7{\cdot}89x.$

Note ....

Thus, we see that if we take deviations from $\bar{x}$ and $\bar{y}$, we get $a = 0$ and we get the value of b directly. In this way the last method is more convenient.

Example 4 : Fit a straight line to the following data.

Year x	:	1951	1961	1971	1981	1991
Production y (000 tons)	:	10	12	8	10	15

Also estimate the production in 1987. **(M.U. 2013)**

Sol. : Now the means of x and y are $\bar{x} = 1971$ and $\bar{y} = 11$.

Since $\bar{x}$ and $\bar{y}$ are round figures, we shall take deviations from 1971 and 11.

Calculations of ΣX, ΣY etc.

x	y	$X = x - 1971$	$Y = y - 11$	X^2	XY
1951	10	– 20	– 1	400	20
1961	12	– 10	1	100	– 10
1971	8	0	– 3	0	0
1981	10	10	– 1	100	– 10
1991	15	20	4	400	80
$N = 5$		0	0	1000	80

Now, $Y = a + bX$

$\therefore \quad \Sigma Y = Na + b\Sigma X$ and $\Sigma XY = a\Sigma X + b\Sigma X^2$

Putting the above values

$0 = 5a + 0 \qquad \therefore \ a = 0$

and $80 = 0 + 1000\, b \qquad \therefore \ b = 0{\cdot}08$

Hence, the equation is $Y = 0{\cdot}08X$.

Putting the values of X, Y, we get

$y - 11 = 0.08\,(x - 1971)$

$\therefore \quad y = - 146{\cdot}68 + 0{\cdot}08x.$

Putting $x = 1987$,

$y = - 146{\cdot}68 + 0{\cdot}08\,(1987) = 12{\cdot}28.$

$\therefore$ The production in 1987 is 12·28.

Example 5 : Fit a straight line of the form $y = a + bx$ to the following data and estimate the value of y for $x = 3{\cdot}5$.

x	:	0	1	2	3	4
y	:	1	1·8	3·3	4·5	6·3

(M.U. 2018)

Sol. : We shall solve this problem by both the methods *i.e.*, by direct method and also by deviation method.

(a) Direct method :

Calculations of Σx, Σx^2 etc.

x	y	x^2	xy
0	1·0	0	0·0
1	1·8	1	1·8
2	3·3	4	6·6
3	4·5	9	13·5
4	6·3	16	25·2
$N = 5$, 10	16·9	30	47·1

Let the equation of the line be $y = a + bx$.

Then the normal equations are

$$\Sigma y = Na + b\Sigma x \quad \therefore \quad 16{\cdot}9 = 5a + 10b \quad \ldots\ldots\ldots\ldots (1)$$

and $$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \therefore \quad 47{\cdot}1 = 10a + 30b \quad \ldots\ldots\ldots\ldots (2)$$

Now, multiply (1) by 2 and subtract the result from (2).

$$\therefore \quad 47{\cdot}1 = 10a + 30b$$

$$33{\cdot}8 = 10a + 20b$$

$$13{\cdot}3 = 10b \qquad \therefore \quad b = 1{\cdot}33$$

Putting this value of b in (1), we get

$$16{\cdot}9 = 5a + 13{\cdot}3 \qquad \therefore \quad 5a = 3{\cdot}6 \qquad \therefore \quad a = 0{\cdot}72$$

Hence, the equation of the line is

$$y = 0{\cdot}72 + 1{\cdot}33x.$$

(b) Deviation method : We shall take deviations of x from the mid-point of the values of x.

x	$X = x - 2$	$Y = y$	X^2	XY
0	− 2	1.0	4	− 2·0
1	− 1	1.8	1	− 1·8
2	0	3.3	0	0·0
3	1	4.5	1	4·5
4	2	6.3	4	12·6
$N = 5$	0	16.9	10	13·3

Let the equation of the line be $Y = a + bX$.

The normal equations are

$$\Sigma Y = Na + b\Sigma X \qquad \text{and} \qquad \Sigma XY = a\Sigma X + b\Sigma X^2$$

Putting the values from the table, we get,

$$16{\cdot}9 = 5a + b\,(0) \qquad \therefore \quad a = \frac{16{\cdot}9}{5} = 3{\cdot}38.$$

And $13{\cdot}3 = a\,(0) + 10\,b \qquad \therefore \quad b = 1{\cdot}33.$

Hence, the equation of the line is $Y = 3{\cdot}38 + 1{\cdot}33X.$

Putting $X = x - 2$ and $Y = y$ we get

$$y = 3{\cdot}38 + 1{\cdot}33\,(x - 2) = 3{\cdot}38 - 2{\cdot}66 + 1{\cdot}33\,x$$

$$\therefore \quad y = 0{\cdot}72 + 1{\cdot}33\,x$$

Thus, the equation of the line is $y = 0{\cdot}72 + 1{\cdot}33x.$

Putting $x = 3{\cdot}5, \quad y = 0{\cdot}72 + 4{\cdot}655 = 5{\cdot}375.$

Hence, when $x = 3{\cdot}5$, $y = 5{\cdot}375.$

Example 6 : Fit a straight line to the following data, with x as independent variable.

x :	1965	1966	1967	1968	1969
y :	125	140	165	195	200

(M.U. 2014)

Sol. : The means of x and y are $\bar{x} = 1967$, $\bar{y} = 165$.

Since $\bar{x}$ and $\bar{y}$ are round figures, we shall take deviations from 1967 and 165.

Calculations of ΣX, ΣY etc.

x	y	$X = x - 1967$	$Y = y - 165$	X^2	XY
1965	125	− 2	− 40	4	80
1966	140	− 1	− 25	1	25
1967	165	0	0	0	0
1968	195	1	30	1	1
1969	200	2	35	4	70
$N = 5$		0	0	10	176

Now, $Y = a + bX$

$\therefore \quad \Sigma Y = Na + b \Sigma X$ and $\Sigma XY = a \Sigma X + b \Sigma X^2$

Putting the above values

$0 = 5a + 0 \qquad \therefore \; a = 0$

and $176 = 0 + 10\, b \qquad \therefore \; b = 17{\cdot}6$

Hence, the equation of line is $Y = 17{\cdot}6X$.

Putting the values of X and Y, we get

$y - 165 = 17{\cdot}6\,(x - 1967)$

$= 165 + 17{\cdot}6\, x - 34619{\cdot}2$

$\therefore \quad y = 17.6x - 34454{\cdot}2.$

Example 7 : Fit a straight line to the following data.

x	:	1	2	3	4	5	6
y	:	83	92	71	90	160	191

Sol. : Now, the mean of $x = 3{\cdot}5$. Since the deviations will be in decimals, we multiply them by 2, *i.e.*, we put $2\,(x - 35) = X$.

The mean of $y = 114{\cdot}5$.

Since, the deviations will be in decimals, we multiply them by 2, *i.e.*, we put $2\,(y - 114{\cdot}5) = Y$.

Calculations of ΣX, ΣY etc.

x	y	$X = 2\,(x - 3{\cdot}5)$	$Y = 2\,(y - 114.5)$	X^2	XY
1	83	− 5	− 63	25	315
2	92	− 3	− 45	9	135
3	71	− 1	− 87	1	87
4	90	1	− 49	1	− 49
5	160	3	91	9	273
6	191	5	153	25	765
$N = 6$		0	0	70	1526

Now, $Y = a + bX$

$\therefore \quad \Sigma Y = Na + b\Sigma X \quad$ and $\quad \Sigma XY = a\Sigma X + b\Sigma X^2$

Putting the above values

$0 = 6a + 0 \qquad \therefore \ a = 0$

and $1526 = 0 + 70\,b \qquad \therefore \ b = 21{\cdot}8$

Hence, the equation of line is $Y = 21{\cdot}8X$.

Putting the values of X and Y, we get

$2\,(y - 114{\cdot}5) = 21{\cdot}8\,(2)\,(x - 3{\cdot}5)$

$\therefore \quad y - 114{\cdot}5 = 21{\cdot}8\,(x - 3{\cdot}5) \qquad \therefore \ y = 38{\cdot}2 + 21{\cdot}8x.$

EXERCISE - V

1. Fit a straight line to the following data :

x :	0	1	2	3	4	5
y :	1	2	3	4·5	6	7·5

[**Ans.** : $y = 0{\cdot}70 + 1{\cdot}32x$]

2. Fit a straight line to the following data :

x :	100	120	140	160	180	200
y :	0·45	0·55	0·60	0·70	0·80	0·85

[**Ans.** : $y = 0{\cdot}041 + 0{\cdot}0041x$]

3. Fit a first degree curve to the following data and estimate the value of y when $x = 73$.

x :	10	20	30	40	50	60	70	80
y :	1	3	5	10	6	4	2	1

[**Ans.** : $y = 4 - 0{\cdot}071u$ where $u = (x - 45) / 5$, $y = 3{\cdot}596$, when $x = 73$]

4. Fit a straight line to the following data and estimate y when $x = 12$

x :	1	2	3	4	5	6	7	8	9	10
y :	52·5	58·7	65·0	70·2	75·4	81·1	87·2	95·5	102·2	108·4

(M.U. 2005) [**Ans.** : $Y = 79{\cdot}62 + 3{\cdot}08X$ where $X = (x - 5{\cdot}5)2$, when $x = 12$, $y = 119{\cdot}66$]

5. Fit a straight line to the following data.

$(x, y) = (-1, -5), (1, 1), (2, 4), (3, 7), (4, 10)$. Estimate y when $x = 7$. **(M.U. 2015)**

[**Ans.** : $y = -2 + 3x$; 19]

18. Fitting a Parabola

Suppose we have a set of values (x_i, y_i). Suppose further that we want to fit a parabola $y = a + bx + cx^2$ to these values. The parabola must be close to the given points as much as possible. The principle of least squares states that the parabola should be such that the distances of the given points from the parabola measured along the y axis must be minimum.

Suppose $P(x_i, y_i)$ is a given point and a line through P parallel to the y axis intersects the curve $y = a + bx + cx^2$ in Q. Then Q is $(x_i, a + bx_i + cx_i^2)$. We find a, b, c such that the distance $|PQ|$ is minimum. But distance $|PQ|$ is minimum when the square of the

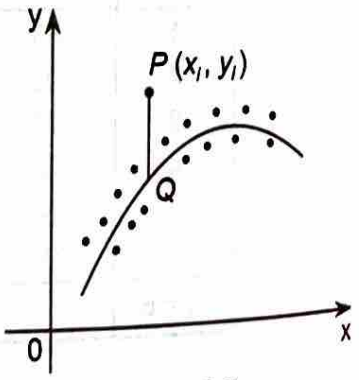


Fig. 2.7

distance $(y_i - a - bx_i - cx_i^2)^2$ is minimum. This must be true for all points. This means we should have $s = \Sigma (y_i - a - bx_i - cx_i^2)^2$ minimum. Since we find a, b, c such that the sum of the squares S is least the method as you know is knowm as the **method of least squares.**

Now for s to be minimum, the conditions are

$$\frac{\partial s}{\partial a} = 0, \quad \frac{\partial s}{\partial b} = 0, \quad \frac{\partial s}{\partial c} = 0$$

Differentiating s partially w.r.t. a, b, c, we get

$$\therefore \quad \Sigma (y_i - a - bx_i - cx_i^2) = 0 ; \qquad \Sigma (y_i - a - bx_i - cx_i^2)\, x_i = 0 ;$$

$$\Sigma (y_i - a - bx_i - cx_i^2)\, x_i^2 = 0.$$

$$i.e., \quad \Sigma y_i = Na + b \Sigma x_i + c \Sigma x_i^2 ; \qquad \Sigma x_i y_i = a \Sigma x_i + b \Sigma x_i^2 + c \Sigma x_i^3 ;$$

$$\Sigma x_i^2 y_i = a \Sigma x_i^2 + b \Sigma x_i^3 + c \Sigma x_i^4.$$

where N is the number of observations.

Dropping the suffix, i, we get the equations as

$$\boxed{\begin{aligned} \Sigma y &= Na + b \Sigma x + c \Sigma x^2 \\ \Sigma xy &= a \Sigma x + b \Sigma x^2 + c \Sigma x^3 \\ \Sigma x^2 y &= a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4 \end{aligned}}$$

These equations are called **normal equations.**

As in the case of straight line for fitting of parabola also we have two methods. (i) Direct Method and (ii) Deviation Method (or Change of Origin Method).

In direct method we use the given values of x and y and find the constants a, b, c.

If the values of x (or y) are evenly spaced we change the origin to the mid-point of x and use the deviations as in Ex. 3 or 4 below.

Example 1 : By the method of least squares find the best values of a, b, c in the law $y = a + bx + cx^2$ to fit the following data.

x :	-2	-1	0	1	2
y :	$-3{\cdot}150$	$-1{\cdot}390$	$0{\cdot}620$	$2{\cdot}880$	$5{\cdot}378$

(M.U. 2019)

Sol. :

Calculations of Σx, Σy, Σx^2

Sr. No.	x	y	x^2	x^3	x^4	xy	x^2y
1	-2	$-3{\cdot}150$	4	-8	16	$6{\cdot}300$	$-12{\cdot}600$
2	-1	$-1{\cdot}390$	1	-1	1	$1{\cdot}390$	$-1{\cdot}390$
3	0	$0{\cdot}620$	0	0	0	$0{\cdot}000$	$0{\cdot}000$
4	1	$2{\cdot}880$	1	1	1	$2{\cdot}880$	$2{\cdot}880$
5	2	$5{\cdot}378$	4	8	16	$10{\cdot}756$	$21{\cdot}512$
$N = 5$	0	$4{\cdot}338$	10	0	34	$21{\cdot}326$	$10{\cdot}402$

The normal equations are

$$\Sigma y = Na + b \Sigma x + c \Sigma x^2 \qquad \therefore \quad 4{\cdot}338 = 5a + 10c$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2 + c \Sigma x^3 \qquad \therefore \quad 21{\cdot}325 = 10b$$

$$\Sigma x^2 y = a \Sigma x^2 + b \Sigma x^3 + c \Sigma x^4 \qquad \therefore \quad 10{\cdot}402 = 10a + 34c$$

Solving these three equations, we get

$$a = 0{\cdot}621,\ b = 2{\cdot}1326,\ c = 0{\cdot}1233$$

Hence, the law is

$$y = 0{\cdot}621 + 2{\cdot}1326x + 0{\cdot}1233x^2.$$

Example 2 : Fit a second degree curve *i.e.*, parabola to the following data.

x :	0	1	2	3	4
y :	1·0	1·8	1·3	2·5	6·3

Sol. : Let the equation of the parabola be $y = a + bx + cx^2$.

Calculations of Σx, Σx^2, etc.

Sr. No.	x	y	x^2	x^3	x^4	xy	x^2y
1	0	1·0	0	0	0	0	0
2	1	1·8	1	1	1	1·8	1·8
3	2	1·3	4	8	16	2·6	5·2
4	3	2·5	9	27	81	7·5	22·5
5	4	6·3	16	64	256	25·2	100·8
$N = 5$	10	12·9	30	100	354	37·1	130·3

The normal equations are

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

Putting the values from the table,

$$12{\cdot}9 = 5a + 10b + 30c \quad \ldots\ldots (1)$$
$$37{\cdot}1 = 10a + 30b + 100c \quad \ldots\ldots (2)$$
$$130{\cdot}3 = 30a + 100b + 354c \quad \ldots\ldots (3)$$

We first eliminate a. For this multiply (1) by 2 and subtract the result from the second.

$$\therefore \quad 37{\cdot}1 = 10a + 30b + 100c$$
$$25{\cdot}8 = 10a + 20b + 60c$$
$$11{\cdot}3 = \quad 10b + 40c \quad \ldots\ldots (4)$$

Now, multiply (1) by 6 and subtract the result from (3),

$$\therefore \quad 130{\cdot}3 = 30a + 100b + 354c$$
$$77{\cdot}4 = 30a + 60b + 180c$$
$$52{\cdot}9 = \quad 40b + 174c \quad \ldots\ldots (5)$$

Now, multiply (4) by 4 and subtract the result from (5).

$$\therefore \quad 52{\cdot}9 = 40b + 174c$$
$$45{\cdot}2 = 40b + 160c$$
$$7{\cdot}7 = \quad 14c$$

$$\therefore \quad c = \frac{7{\cdot}7}{14} = 0{\cdot}55.$$

Putting this value of c in (4), we get,

$$11{\cdot}3 = 10b + 40\ (0{\cdot}55)$$

$$\therefore\ 10b = 11{\cdot}3 - 22 = -10{\cdot}7 \qquad \therefore\ b = -1{\cdot}07$$

Putting the values of b and c in (1), we get

$$12{\cdot}9 = 5a + 10\ (-1{\cdot}07) + 30\ (0{\cdot}55)$$

$$\therefore\ 5a = 12{\cdot}9 + 10{\cdot}7 - 16{\cdot}5 = 7{\cdot}1 \qquad \therefore\ a = 1{\cdot}42$$

Hence, the equation of the parabola is

$$y = 1{\cdot}42 - 1{\cdot}07x + 0{\cdot}55x^2.$$

Example 3 : Fit a non-linear trend of the form $y = a + bx + cx^2$ to the following data.

x :	0	1	2	3	4
y :	1·0	1·5	1·5	2·5	3·5

Sol. : **Calculations of $\Sigma X, \Sigma Y, \Sigma X^2$**

Sr. No.	x	$Y = y$	Deviations from mid point $X = x - 2$	X^2	X^3	X^4	XY	X^2Y
1	0	1·0	−2	4	−8	16	−2·0	4·0
2	1	1·5	−1	1	−1	1	−1·5	1·5
3	2	1·5	0	0	0	0	0·0	0·0
4	3	2·5	1	1	1	1	2·5	2·5
5	4	3·5	2	4	8	16	7·0	14·0
$N = 5$	10	10·0	0	10	0	34	6·0	22·0

Thus, we have

$$N = 5,\ \Sigma X = 0,\ \Sigma Y = 10,\ \Sigma X^2 = 10,\ \Sigma X^3 = 0,\ \Sigma X^4 = 34,\ \Sigma XY = 6,\ \Sigma X^2 Y = 22.$$

Since, we have taken mid-point of X as origin $\Sigma X = 0$ and $\Sigma X^3 = 0$. The normal equations now reduce to

$$\Sigma Y = Na + b\Sigma X^2$$

$$\Sigma XY = a\Sigma X^2$$

$$\Sigma X^2 Y = a\Sigma X^2 + c\Sigma X^4$$

Putting the above values, we get $10 = 5a + 10c$

$$\therefore \quad 22 = 10a + 34c \quad \text{and} \quad 6 = 10b \qquad \therefore\ b = \frac{6}{10}$$

$$20 = 10a + 20c$$

$$2 = \quad 14c \qquad \therefore\ c = \frac{2}{14} = \frac{1}{7}$$

$$\therefore\ 10 = 5a + \frac{10}{7} \qquad \therefore\ a = \frac{12}{7}. \qquad \therefore \text{ The equation is } Y = \frac{12}{7} + \frac{6}{10}X + \frac{1}{7}X^2.$$

Now putting $Y = y$ and $X = x - 2$, we get,

$$y = \frac{12}{7} + \frac{6}{10}(x-2) + \frac{1}{7}(x-2)^2 = \frac{12}{7} + \frac{6}{10}x - \frac{12}{10} + \frac{1}{7}x^2 - \frac{4}{7}x + \frac{4}{7}$$

$$\therefore\ y = \frac{38}{35} + \frac{1}{35}x + \frac{1}{7}x^2.$$

Example 4 : Fit a second degree parabolic curve to the following data.

x : 1, 2, 3, 4, 5, 6, 7, 8, 9.

y : 2, 6, 7, 8, 10, 11, 11, 10, 9. **(M.U. 2004, 07, 12)**

Sol. : Since the values of x are odd and are equally spaced we change x to X by the relation $X = x - 5$ and put $y = Y$.

Let the equation of the parabola be $Y = a + bX + cX^2$. Then the normal equations are

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Calculations of ΣX, ΣX^2 etc.

Sr. No.	x	$X = x - 5$	$Y = y$	X^2	X^3	X^4	XY	X^2Y
1	1	−4	2	16	−64	256	−8	32
2	2	−3	6	9	−27	81	−18	54
3	3	−2	7	4	−8	16	−14	28
4	4	−1	8	1	−1	1	−8	8
5	5	0	10	0	0	0	0	0
6	6	1	11	1	1	1	11	11
7	7	2	11	4	8	16	22	44
8	8	3	10	9	27	81	30	90
9	9	4	9	16	64	256	36	144
$N = 9$		0	74	60	0	708	51	411

Now the normal equations become

$$74 = 9a + 60c \quad \text{..... (1)}$$

$$51 = 60b \quad \therefore b = 0{\cdot}85 \quad \text{..... (2)}$$

$$411 = 60a + 708c \quad \text{..... (3)}$$

Now to find a and c multiply (1) by 60 and (3) by 9.

$$\therefore \quad 4440 = 540a + 3600c$$

$$3699 = 540a + 6372c$$

$$741 = -2772c \quad \therefore c = -0{\cdot}2673$$

Now, $9a = 74 - 60c = 74 + 16{\cdot}038 = 90{\cdot}038$

$\therefore a = 10{\cdot}0042 = 10$

$\therefore$ The equation of the parabola is $Y = 10 + 0{\cdot}85X - 0{\cdot}27X^2$.

i.e. $y = 10 + 0{\cdot}85\,(x - 5) - 0{\cdot}27\,(x - 5)^2 \quad \therefore y = -1 + 3{\cdot}55x - 0{\cdot}27x^2$.

Example 5 : Fit a second degree parabolic curve to the following data and estimate the production in 1982.

Year (x) :	1974	75	76	77	78	79	80	91
Production (y) (In tons) :	12	14	26	42	40	50	52	53

Sol. : Since the values of X though not odd are equispaced we change x to X by the relation $X = (x - 1977{\cdot}5)2$ and we put $y = Y$.

Let the equation of the parabola be $Y = a + bX + cX^2$. Then the normal equations are

$$\Sigma Y = Na + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Calculations of ΣX, ΣX^2 etc.

Sr. No.	x	X	Y	X^2	X^3	X^4	XY	X^2Y
1	1974	− 7	12	49	− 343	2401	− 84	588
2	1975	− 5	14	25	− 125	625	− 70	350
3	1976	− 3	26	9	− 27	81	− 78	234
4	1977	− 1	42	1	− 1	1	− 42	42
5	1978	1	40	1	1	1	40	40
6	1979	3	50	9	27	81	150	450
7	1980	5	52	25	125	625	260	1300
8	1981	7	53	49	343	2401	371	2597
$N = 9$		0	289	168	0	6216	547	5601

Now, the normal equations become

$289 = 8a + 168c$ ..... (1)

$547 = 168b$ $\therefore$ $b = 3{\cdot}2559$ ..... (2)

$5601 = 168a + 6216c$ ..... (3)

Multiply (1) by 21 and subtract (3) from the result,

$6069 = 168a + 3528c$

$5601 = 168a + 6216c$

$468 = -2688c$

Now $289 = 8a + 168c$ $\therefore$ $c = -0{\cdot}1741$

$\therefore$ $289 = 8a + (168)(-0{\cdot}1741)$ $\therefore$ $a = 39{\cdot}7811$

The equation to the curve is $Y = 39{\cdot}78 + 3{\cdot}26X - 0{\cdot}17X^2$.

Putting $Y = y$ and $X = (x - 1977{\cdot}5)2$, the equation of the parabola is

$y = 39{\cdot}78 + 3{\cdot}26\,(x - 1977{\cdot}5)2 - 0{\cdot}17\,(x - 1977{\cdot}5)^2 \cdot 4$

$y = -2671997{\cdot}77 + 2695{\cdot}92x - 0{\cdot}68x^2$.

Putting $x = 1982$, we get $y = 55{\cdot}35$.

EXERCISE - VI

1. Fit a parabola to the following data and estimate the valu

x :	1	2	3	4	5
y :	25	28			

2. Fit a second parabola to the following data taking x as the independent variable and shifting the origin to 2 for x.

x :	0	1	2	3	4
y :	1	1·8	1·3	2·5	6·3

Find the difference between the actual value and estimated value of y when $x = 2$.

[**Ans.** : $Y = 1·48 + 1·13X + 0·55X^2$ where $X = x - 2$. Difference $= -0·18$]

3. Fit a parabola to the following data :

x :	-2	-1	0	1	2
y :	1·0	1·8	1·3	2·5	6·3

(M.U. 2006) [**Ans.** : $y = 1·48 + 1·13x + 0·55x^2$]

4. Fit a parabola to the following data :

x :	-2	-1	0	1	2
y :	$-3·150$	$-1·390$	0·620	2·880	5·378

[**Ans.** : $y = 0·621 + 2·1326x + 0·1233x^2$]

5. Fit a second degree curve to the following data and estimate the value of y when $x = 80$

x :	10	20	30	40	50	60	70
y :	20	60	70	80	90	100	100

[**Ans.** : Put $X = (x - 40) / 10$, $Y = y / 10$, $Y = 8·381 + 1·2143X - 0·2381X^2$ when $x = 80$, $y = 94·286$]

6. Fit a parabola to the following data and estimate y when $x = 10$.

x :	1	2	3	4	5	6	7	8	9
y :	2	6	7	8	10	11	11	10	9

(M.U. 2004)

[**Ans.** : $Y = 3 + 0·85X - 0·27X^2$ where $X = x - 5$, $Y = y - 7$, $y = 7·5$]

7. Fit a second degree curve to the following data and estimate the production in 1975.

Year :	1921,	31,	41,	51,	61,	71,	81.
Production (in tons) :	3,	5,	9,	10,	12,	14,	15.

[**Ans.** : $Y = 10·33 + 2·0357X - 0·1547X^2$ where $X = (x - 1951) / 10$. Production = 14·32]

8. Fit a second degree curve to the following data

Year :	1965,	66,	67,	68,	69,	70,	71,	72.
Profit (in Crores Rs.) :	125,	140,	165,	195,	200,	215,	220,	230.

Also estimate the profit in 1973.

[**Ans.** : $Y = 194·68 + 7·68X - 0·40X^2$ where $X = (x - 1968·5) / 2$. And in 1973 profit = 231·4 crores of Rs.]

EXERCISE - VII

Theory : Correlation

1. What is meant by correlation ? Describe scatter diagram and interpret.

2. Define - (i) Karl Pearson's coefficient of correlation, (ii) Spearman's rank correlation coefficient, (iii) Coefficient of variation. **(M.U. 2005)**

3. Two variables x and y are connected by the relation $ax + by + c = 0$. Show that the coefficient of correlation is either + 1 or – 1.

4. Define product moment correlation coefficient and show that it is always numerically less than or equal to unity.

5. Define Karl Pearson's Product moment correlation coefficient.

6. Prove that Spearson's rank correlation coefficient R is given by

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

(M.U. 1996, 99, 2002, 03, 04, 05, 06, 09)

7. What is scatter diagram ? How does it help in studying the correlation between two variables. Draw scatter diagrams for $r = +1$, $r = -1$ and $r = 0$.

8. Define the Karl Pearson's coefficient of correlation r between two variables x and y. What is "Spurious Correlation" ?

Interpret the cases $r = +1$, $r = -1$, $r = 0$. Also draw the scatter diagrams corresponding to these cases.

Theory : Regression

1. Distinguish between correlation and regression.

2. Explain "the line of regression". Why there are two lines of regression ? **(M.U. 2007)**

3. Explain what you understand by regression. What are lines of regression ? Why are there in general two lines of regression ? When do they coincide, when are they perpendicular ? **(M.U. 2004)**

4. Explain the method of scatter diagram to obtain a line of regression.

5. Obtain the equations of lines of regression. **(M.U. 2002, 03)**

6. Prove that the sum of the coefficients of regression is greater than or equal to $2r$ where r is the coefficient of correlation. [See 3, page 2-31] **(M.U. 1998)**

7. Find the expression for the acute angle between the lines of regression. **(M.U. 2004, 05)**

8. With usual notation prove that **(i)** $r = \sqrt{b_{yx} \cdot b_{xy}}$, **(ii)** $b_{xy} + b_{yx} \geq 2r$.

9. Examine whether the following statement is correct

$b_{xy} = 3{\cdot}2$, $b_{yx} = 0{\cdot}7$. **[Ans. : No]**

10. If θ is the angle between the two lines of regression, prove that

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

(M.U. 1996, 2004, 07)

☆☆☆