$$= 1 - \frac{54}{504}$$

$$= 1 - 0.1071$$

$$= 0.8929$$

$$\therefore \quad r \simeq 0.89$$

The value of r is near to 1. So, it can be said that there is a high degree of positive correlation between the sales and profit.

Note :

(1) The sum of difference in the ranks R_x and R_y is always zero. i.e. $\Sigma d = \Sigma (R_x - R_y) = 0$

(2) If $R_x = R_y$ for each pairs of the observations of two variables x and y then all corresponding values of d will be zero and hence $\Sigma d^2 = 0$. In this case, the value of r will be 1.

(3) If the ranks R_x and R_y are in exact reverse order of each other (see illustration 18) then $r = -1$.

Activity

Collect the information regarding the marks obtained by any ten students of your class in the subjects of Statistics and Economics. Find the correlation coefficient between the marks of two subjects using Karl Pearson's and Spearman's method and compare them.

Illustration 24 : A transport company wants to know the relation between driving experience and the number of accidents by the drivers. The sum of squares of differences in the ranks given to driving experience and the number of accidents by eight drivers is found to be 126. Find the rank correlation coefficient.

Here, $n = 8$ and the sum of squares of difference in the ranks is 126, i.e. $\Sigma d^2 = 126$.

$$r = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(126)}{8(64 - 1)}$$

$$= 1 - \frac{756}{504}$$

$$= 1 - 1.5$$

$$r = -0.5$$

Illustration 25 : Ten students selected from various schools of a district were ranked on the basis of their proficiency in Sports and General knowledge. The rank correlation coefficient obtained from the data was found to be 0.2. Later on, it was noticed that the difference in the ranks of the two attributes for one of the students was taken as 3 instead of 2. Find the correct value of rank correlation coefficient.

Here, $n = 10$

Incorrect $d = 3$

Correct $d = 2$

$$\text{Now,} \quad r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$\therefore \quad 0.2 = 1 - \frac{6\Sigma d^2}{10(100-1)}$$

$$\therefore \quad 0.2 = 1 - \frac{6\Sigma d^2}{990}$$

$$\therefore \quad \frac{6\Sigma d^2}{990} = 1 - 0.2$$

$$\therefore \quad \frac{6\Sigma d^2}{990} = 0.8$$

$$\therefore \quad \Sigma d^2 = \frac{0.8 \times 990}{6}$$

$$\therefore \quad \Sigma d^2 = 132$$

Since one difference 2 is wrongly taken as 3, the corrected value of Σd^2 is obtained as follows :

$$\text{Corrected } \Sigma d^2 = 132 - (\text{Wrong } d)^2 + (\text{Correct } d)^2$$

$$= 132 - 3^2 + 2^2$$

$$= 132 - 9 + 4$$

$$= 127$$

$\therefore$ Correct value of the rank correlation coefficient is obtained as follows :

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(127)}{10(100-1)}$$

$$= 1 - \frac{762}{990}$$

$$= 1 - 0.7697$$

$$= 0.2303$$

$$\therefore r \simeq 0.23$$

Merits and Limitations of Spearman's Rank Correlation Method

Merits :

(1) This method is easy to understand.

(2) The calculation of rank correlation method is easier than that of Karl Pearson's correlation coefficient.

(3) It is the only method when the data is qualitative.

(4) When dispersion is more or when the extreme observations are present in the data, Spearman's formula is preferred over Karl Pearson's formula.

Limitations :

(1) Since the ranks are used instead of the actual observations, there is always a loss of some information. So, this method does not provide accurate value of the correlation coefficient as compared to Karl Pearson's method.

(2) Unless the ranks are given, it is tedious to assign ranks when the number of observations is large.

(3) This method can not be used for a bivariate frequeny distribution. (In such a case, Karl Pearson's method is used and you will learn it in your higher studies.)

Exercise 2.3

1. Six companies are ranked by the two market analysts on the basis of their growth in the recent past.

Company	A	B	C	D	E	F
Rank by Analyst 1	5	2	1	4	3	6
Rank by Analyst 2	6	4	3	2	1	5

Find the rank correlation coefficient between the evaluation given by two analysts.

2. An official has ranked nine villages of a sample on the basis of the work done in the area of 'Swachhata Abhiyan' and 'Beti Bachavo Abhiyan' by the villages. The ranks are given below.

Village	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Rank for Swachhata Abhiyan	4	8	7	1	9	5	6	2	3
Rank for Beti Bachavo Abhiyan	6	8	5	1	9	7	3	4	2

Find the rank correlation coefficient between the performances of the villages in two Abhiyans.

3. The following information is obtained by a survey conducted by a town planning committee of a state.

City	A	B	C	D	E
Population (lakh)	57	45	14	18	8
Rate of growth (per thousand)	13	20	10	15	5

Find the rank correlation coefficient between the population of the cities and the rate of growth of the population.

4. The following information is obtained by taking a sample of ten students from the students of a Science college.

Student	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics	39	65	62	90	82	75	25	98	36	78
Marks in Statistics	47	53	58	86	62	68	60	91	51	84

Find the rank correlation coefficient between the ability of the students in the subjects of Mathematics and Statistics.

5. From the following information of heights of husband and wife, calculate the rank correlation coefficient between their heights.

Height of husband (cms)	156	153	185	157	163	191	162
Height of wife (cms)	154	148	162	157	162	170	154

6. Two interviewers gave the following scores to the candidates on the basis of their performance in the interview. Find the rank correlation coefficient between the evaluation of two interviewers.

Candidate	A	B	C	D	E	F	G	H
Marks by first interviewer	28	44	10	28	47	35	19	40
Marks by second interviewer	32	45	25	32	41	32	24	38

7. Ten contestants are ranked in a beauty contest by two judges and the sum of squares of differences in their ranks is found to be 214. Find the rank correlation coefficient.

8. The coefficient of rank correlation of the marks obtained by 10 students in two particular subjects was found to be 0.5. Later on, it was found that one of the differences of the ranks of a student was 7 but it was taken as 3. Find the corrected value of the correlation coefficient.

*

2.9 Precautions in the Interpretation of Correlation Coefficient

The coefficient of correlation measures the strength of linear relationship between two variables. An erroneous interpretation of r may lead us to a misunderstanding about the relationship between two variables. The following are some of the points to be kept in mind as a precaution :

(1) Correlation is only a measure of strength of linear relationship between two variables. It gives no indication about presence of cause and effect relationship between them and it does not give any idea about the information that out of the two, which variable is the dependent (effect) and the other as independent (cause). The interpretation of the correlation coefficient depends very much on experience. The investigator must have thorough knowledge about the variables under consideration and the various factors which affect these variables. Several examples can be cited indicating no meaningful correlation between two variables though the value of $|r|$ is very near to 1. Generally, it happens when r is calculated without prior knowledge about cause and effect relationship between the variables. For example, the two series of data relating to the number of persons died in road accidents in a city and the price of Tuver Dal during the same period may exhibit a high correlation (i.e. r may be near to 1). But there can not be meaningful relationship between them. Therefore, this kind of correlation is known as nonsense or spurious correlation.

(2) Sometimes, due to the presence of other factors, the value of $|r|$ between given two variables may be close to 1 though two variables are not correlated. For example, the data relating to the yield of rice and sugarcane show a fairly high degree of positive correlation though there is no connection between these two variables. This may be due to the favourable effect of external factors like weather conditions, irrigation system, fertilizers etc.

(3) When $r = 0$, we can merely say that there is no linear correlation. i.e. there is a lack of linear correlation. But there may be a non linear (quadratic or any other type) relationship between the variables. e.g. :

x	–4	–3	–2	–1	1	2	3	4
y	16	9	4	1	1	4	9	16

If we calculate the Karl Pearson's coefficient of correlation for the above example then the value of r will be 0. So, we may interpret that the two variables are uncorrelated but it is a wrong interpretation. If we observe the values of two variables X and Y then we can see that they have the relation $Y = X^2$. This relation is not linear but it is quadratic. So, though there is a perfect quadratic relationship between the two variables, we get $r = 0$. So, from this example we can understand that $r = 0$ suggests a lack of linear correlation only but there may be other kind of correlation.

(4) If the correlation coefficient computed from bivariate data which is related to a given region or class or given time duration then its interpretation should be limited to that region or class or time duration only. The interpretation of r computed from such data should not be extended or generalised outside the region or class or time duration without proper verification in order to avoid any kind of misunderstanding.

e.g. If a company starts manufacturing a new product and advertises it for its sale then initially by increasing the advertisement cost, sale of the product also increases when the quality of product is good. But after some time limit, sale of the product may not increase even if its advertisement cost increases. Normally there is high degree of positive correlation between the advertisement cost and sales. During initial production period. But after some time that may not be the case. So, the interpretation that there is a high degree of positive correlation between the advertisement cost and sales can not be applied for the data outside its time period.

Some Illustrations :

Illustration 26 : Determine the value of the correlation coefficient from the following results.

$Cov(x, y) : s_x^2 = 3 : 5$ **and** $s_x : s_y = 1 : 2$

Here, $Cov(x, y) : s_x^2 = 3 : 5 \quad \therefore \quad \frac{Cov(x, y)}{s_x^2} = \frac{3}{5}$

and $s_x : s_y = 1 : 2 \quad \therefore \quad \frac{s_x}{s_y} = \frac{1}{2}$

Now, $r = \frac{Cov(x, y)}{s_x \, s_y} = \frac{Cov(x, y)}{s_x^2} \times \frac{s_x}{s_y}$

$= \frac{3}{5} \times \frac{1}{2}$

$= \frac{3}{10}$

$\therefore \quad r = 0.3$

Illustration 27 : The following results are obtained from a bivariate data.

n = 10, $\Sigma(x-\bar{x})(y-\bar{y}) = 72$, $s_x = 3$ and $\Sigma(y-\bar{y})^2 = 160$ Find the correlation coefficient.

From the available results, first we shall find s_y.

$$s_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n}} = \sqrt{\frac{160}{10}} = \sqrt{16} = 4$$

Now, substituting the necessary values in the following formula,

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{ns_x s_y}$$

$$= \frac{72}{10(3)(4)}$$

$$= \frac{72}{120}$$

$$\therefore \quad r = 0.6$$

Illustration 28 : An educationalist has conducted an experiment to know the relation between the usage of Social Media in mobile phone and the result of the examination. A group of 10 students is selected for this and the following results were obtained regarding, the time spent x (in hours) in last week on Social Media and the marks (y) obtained out of 50 in the examination, taken immediately after it.

$\Sigma x = 133$, $\Sigma y = 220$, $\Sigma x^2 = 2344$, $\Sigma y^2 = 6500$ and $\Sigma xy = 3500$

Later on, it was found that one of the pairs of observations of X and Y was taken as (13, 20) instead of (15, 25). Find the correct value of the correlation coefficient between X and Y.

Here, n = 10, $\Sigma x = 133$, $\Sigma y = 220$, $\Sigma x^2 = 2344$, $\Sigma y^2 = 6500$ and $\Sigma xy = 3500$

Incorrect Pair : (13, 20)

Correct Pair : (15, 25)

Now, we find corrected values of these measures as follows :

$\Sigma x = 133 - 13 + 15 = 135$

$\Sigma x = 220 - 20 + 25 = 225$

$\Sigma x^2 = 2344 - (13)^2 + (15)^2 = 2344 - 169 + 225 = 2400$

$\Sigma y^2 = 6500 - (20)^2 + (25)^2 = 6500 - 400 + 625 = 6725$

$\Sigma xy = 3500 - (13 \times 20) + (15 \times 25) = 3500 - 260 + 375 = 3615$

Substituting these corrected values in the following formula,

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

$$= \frac{10(3615) - (135)(225)}{\sqrt{10(2400) - (135)^2} \cdot \sqrt{10(6725) - (225)^2}}$$

$$= \frac{36150 - 30375}{\sqrt{24000 - 18225} \cdot \sqrt{67250 - 50625}}$$

$$= \frac{5775}{\sqrt{5775} \cdot \sqrt{16625}}$$

$$= \frac{5775}{\sqrt{96009375}}$$

$$= \frac{5775}{9798.4374}$$

$$= 0.5894$$

$$\therefore \quad r \simeq 0.59$$

Illustration 29 : (1) If the correlation coefficient between two variables X and Y is 0.5, find the value of the following : (i) $r(x, -y)$ (ii) $r(-x, y)$ (iii) $r(-x, -y)$

Here, $r(x, y) = 0.5$

From the property no. 5 of correlation coefficient,

(i) $r(x, -y) = -r(x, y) = -0.5$

(ii) $r(-x, y) = -r(x, y) = -0.5$

(iii) $r(-x, -y) = r(x, y) = 0.5$

(2) If $r(x, y) = 0.8$ then find $r(u, v)$ for the following.

(i) $u = x - 10$ and $v = y + 10$

(ii) $u = \frac{x-5}{3}$ and $v = 2y + 7$

(iii) $u = \frac{2x-3}{10}$ and $v = \frac{10-y}{100}$

(iv) $u = \frac{5-x}{2}$ and $v = \frac{5+y}{2}$

(v) $u = \frac{20-x}{3}$ and $v = \frac{20-y}{7}$

While defining u and v from the properties (no. 4 and no. 5), the value of $r(u, v)$ will be dependent on the signs of the coefficients of X and Y.

i.e. $r(u, v) = r(x, y)$ or $-r(x, y)$

(i) $r(x - 10, y + 10) = r(u, v) = 0.8$

(ii) $r\left(\frac{x-5}{3}, 2y + 7\right) = r(u, v) = 0.8$

(iii) $r\left(\frac{2x-3}{10}, \frac{10-y}{100}\right) = r(u, v) = -0.8$

(iv) $r\left(\frac{5-x}{2}, \frac{5+y}{2}\right) = r(u, v) = -0.8$

(v) $r\left(\frac{20-x}{3}, \frac{20-y}{7}\right) = r(u, v) = 0.8$

Illustration 30 : A project is conducted by the group of the students of an MBA Institute to know the relation between the results of the final year of school and final year of graduation for the students. The following information is obtained from a sample of 10 students regarding the percentage of marks in standard 12 (x) and the percentage of marks in the final year of graduation (y).

$$n = 10, \Sigma(x-65) = -2, \ \Sigma(y-60) = 2, \ \Sigma(x-65)^2 = 176, \ \Sigma(y-60)^2 = 140, \Sigma(x-65)(y-60) = 141$$

Find the correlation coefficient between the percentages of marks in Standard 12 and the final year of graduation.

Here $\Sigma(x-65) = -2 \neq 0$ $\quad\therefore\ A = 65$

$\Sigma(y-60) = 2 \neq 0$ $\quad\therefore\ B = 60$

(Here, the sum of deviations are not zero, so $65 \neq \bar{x}$ and $60 \neq \bar{y}$)

Now, let us define $u = (x-65)$ and $v = (y-60)$.

So, $\Sigma(x-65) = \Sigma u = -2, \ \Sigma(y-60) = \Sigma v = 2$

$\Sigma(x-65)^2 = \Sigma u^2 = 176, \ \Sigma(y-60)^2 = \Sigma v^2 = 140$

$\Sigma(x-65)(y-60) = \Sigma uv = 141$

Substituting the above values in the following formula,

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10(141) - (-2)(2)}{\sqrt{10(176) - (-2)^2} \cdot \sqrt{10(140) - (2)^2}}$$

$$= \frac{1414}{\sqrt{1756} \cdot \sqrt{1396}}$$

$$= \frac{1414}{\sqrt{2451376}}$$

$$= \frac{1414}{1565.6871}$$

$$= 0.9031$$

$$\therefore \quad r \simeq 0.90$$

Illustration 31 : To study the relation between the age (X years) of teenage children and their daily requirement of protein (y grams), the following information is obtained from a sample of 10 children taken by the Health Department of State.

$$\Sigma x = 140, \ \Sigma y = 150, \ \Sigma(x-10)^2 = 180, \quad \Sigma(y-15)^2 = 215, \Sigma(x-10)(y-15) = 60$$

Find the correlation coefficient between X and Y.

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{140}{10} = 14, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{150}{10} = 15$

We can see that the deviations are not taken from actual mean $(\bar{x} = 14)$ for the variable X. So, to solve the example, it will be convenient to define $u = (x - A) = (x - 10)$ and $v = (y - B) = (y - 15)$.

We are given the following information.

$\Sigma(x-10)^2 = \Sigma u^2 = 180,\ \Sigma(y-15)^2 = \Sigma v^2 = 215,\ \Sigma(x-10)(y-15) = \Sigma uv = 60$

Now, in order to use an appropriate formula of r, first we need Σu and Σv.

$\Sigma u = \Sigma(x-10) = \Sigma x - \Sigma 10 = \Sigma x - n(10) = 140 - 10(10) = 140 - 100 = 40$

$\Sigma v = \Sigma(y-15) = \Sigma y - \Sigma 15 = \Sigma y - n(15) = 150 - 10(15) = 150 - 150 = 0$

$\{\because\ \Sigma k = \underbrace{k + k + k + \ldots\ldots\ldots + k}_{n \text{ times}} = nk$ where, k = constant$\}$

Substituting the above values in the following formula,

$$r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{10(60) - (40)(0)}{\sqrt{10(180) - (40)^2} \cdot \sqrt{10(215) - (0)^2}}$$

$$= \frac{600 - 0}{\sqrt{1800 - 1600} \cdot \sqrt{2150 - 0}}$$

$$= \frac{600}{\sqrt{200} \cdot \sqrt{2150}}$$

$$= \frac{600}{\sqrt{430000}}$$

$$= \frac{600}{655.7439}$$

$$= 0.9150$$

$\therefore\ r \simeq 0.92$

Illustration 32 : To know the relation between the ability in two different subjects for the students, a sample of seven students is taken from a school. From the information of marks in two subjects for 7 students, it is known that the sum of the squares of differences in the ranks of these marks is 25.5. It is also known that two students got equal marks in one subject and all the remaining marks are different. Find the rank correlation coefficient.

Here, $n = 7$ and $\Sigma d^2 = 25.5$

Two students got equal marks in a subject $(\therefore m = 2)$. So, we can say that there is a tie in assigning the ranks. Therefore, we need to take the term $\left(\frac{m^3 - m}{12}\right)$ only once to obtain CF.

$$CF = \left(\frac{m^3-m}{12}\right) = \left(\frac{2^3-2}{12}\right) = 0.5$$

$$r = 1 - \frac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$$

$$= 1 - \frac{6\left[25.5 + 0.5\right]}{7(49-1)}$$

$$= 1 - \frac{6(26)}{336}$$

$$= 1 - \frac{156}{336}$$

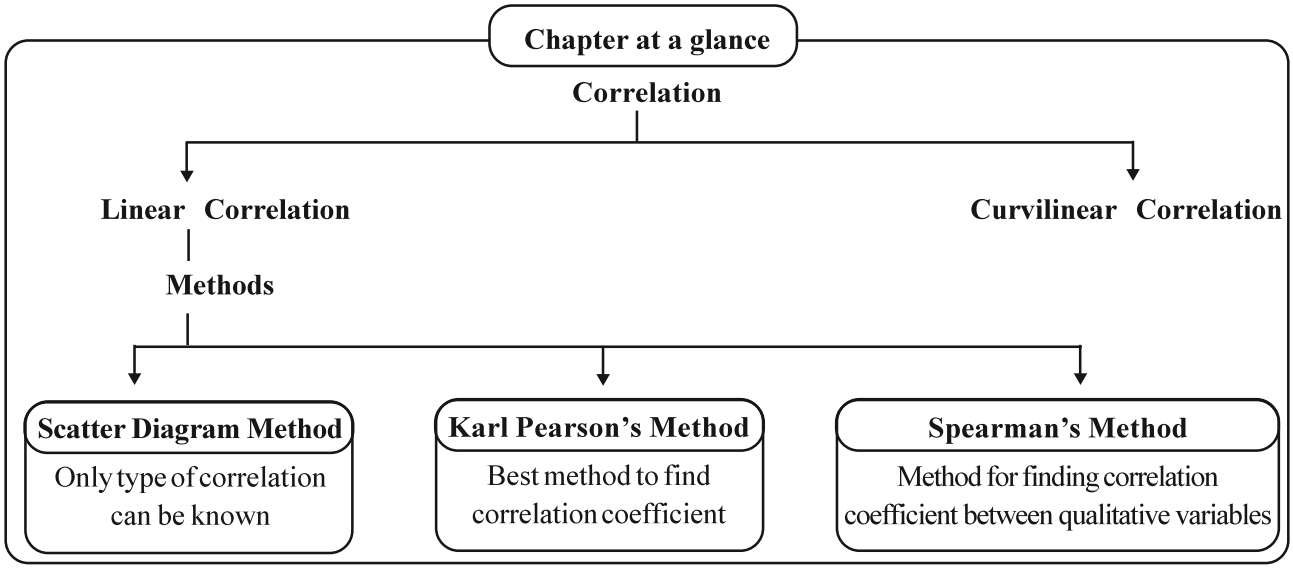
$$= 1 - 0.4643$$

$$= 0.5357$$

$$\therefore \quad r \simeq 0.54$$

Summary

- **Correlation :** Simultaneous change in the values of two variables and direct or indirect cause-effect relationship between them.
- **Linear Correlation :** There are almost constant proportional changes in the values of two variables i.e. the points corresponding to the values of two correlated variables are on or nearer to a line.
- **Positive Correlation :** The changes in the values of two correlated variables are in the same direction.
- **Negative Correlation :** The changes in the values of two correlated variables are in the opposite direction.
- **Correlation Coefficient :** The numerical measure showing the strength of linear correlation between two variables is a correlation coefficient.
- **Scatter diagram :** A simple method for identifying linear correlation and its type (positive or negative).
- **Karl Pearson's Method :** The best method of obtaining type and strength of linear correlation using all observations.
- **Spearman's Rank Correlation Method :** A method for obtaining the correlation coefficient for qualitative variables and also preferable when dispersion is more in quantitative variables.
- The cause and effect relation between two variables cannot be proved but under the assumption that it does exist, the concept of correlation is studied.
- $r = 0$ indicates the absence of linear correlation only but there may be other type of correlation.



List of Formulae

Karl Pearson's Method :

Correlation coefficient $= r$

(1) $r = \dfrac{\text{Covariance}}{(\text{S.D of } X)(\text{S.D of } Y)} = \dfrac{Cov(X,Y)}{s_x \cdot s_y}$

Where, $Cov(X,Y) = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \dfrac{\Sigma xy - n\bar{x}\,\bar{y}}{n}$

$s_x = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n}}$ and $s_y = \sqrt{\dfrac{\Sigma(y-\bar{y})^2}{n}}$

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

(3) $r = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

(4) $r = \dfrac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \cdot \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$ Where, $u = x - A$ or $\dfrac{x-A}{c_x}$, $v = y - B$ or $\dfrac{y-B}{c_y}$

(5) $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n \cdot s_x \cdot s_y}$

(6) $r = \dfrac{\Sigma xy - n\bar{x}\,\bar{y}}{n \cdot s_x \cdot s_y}$

Specially for short sums

Spearman's Rank Correlation Method

(7) $r = 1 - \dfrac{6\Sigma d^2}{n(n^2-1)}$ When the observations are not repeated

(8) $r = 1 - \dfrac{6\left[\Sigma d^2 + CF\right]}{n(n^2-1)}$ When some of the observations are repeated

Where, d = Rank of x – Rank of y = $R_x - R_y$

CF = Correction Factor = $\Sigma\left(\dfrac{m^3-m}{12}\right)$

m = Number of times an observation is repeated

Exercise 2

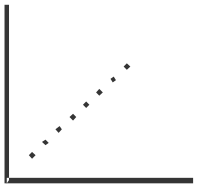
Section A

Find the correct option for the following multiple choice questions :

1. In context with correlation, what do you call the graph, if the points of paired observations (x,y) are shown in a graph ?

(a) Histogram (b) Circle diagram (c) Scatter diagram (d) Frequency curve

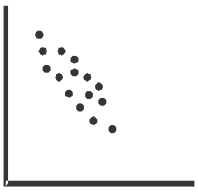
2. Which kind of the correlation exists if the following scatter diagram is of two variables X and Y ?



(a) Perfect Positive correlation (b) Partial Positive correlation

(c) Perfect Negative correlation (d) Partial Negative correlation

3. Which kind of the correlation exists if the following scatter diagram is of two variables X and Y ?



(a) Perfect Positive correlation (b) Partial Positive correlation

(c) Perfect Negative correlation (d) Partial Negative correlation

4. What is the value of r, if all the points plotted in a scatter diagram lie on a single line only ?

(a) 0 (b) 1 or -1 (c) 0.5 (d) -0.5

5. What is the range of the correlation coefficient r ?

(a) $-1 < r < 1$ (b) 0 to 1 (c) $-1 \leq r \leq 1$ (d) -1 to 0

6. The measurement unit of a variable 'Weight' is kg. and that of 'Height' is cm. What can you say about the measurement unit of the correlation coefficient between them ?

(a) kg (b) cm (c) km (d) does not have any unit

7. Which kind of the correlation can be obtained if the two variables are varying in opposite direction in constant proportion ?

(a) Partial Positive Correlation (b) Perfect Negative Correlation

(c) Perfect Positive Correlation (d) Partial Negative Correlation

8. What does the numerator indicate in the formula for calculating the correlation coefficient by Karl Perason's method ?

(a) Product of variance of X and Y (b) Covariance of X and Y

(c) Variance of X (d) Variance of Y

9. Which of the following values is not possible as a value of r ?

(a) 0.99 (b) -1.07 (c) -0.85 (d) 0

10. If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$, $c_x > 0$, $c_y > 0$ then which of the following statement is correct ?

(a) $r(x, y) \neq r(u, v)$ (b) $r(x, y) > r(u, v)$ (c) $r(x, y) = r(u, v)$ (d) $r(x, y) < r(u, v)$

11. If $r(x, y) = 0.7$ then what is the value of $r(x + 0.2, y + 0.2)$?

(a) 0.7 (b) 0.9 (c) 1.1 (d) –0.7

12. If $r(-x, y) = -0.5$ then what is the value of $r(x, -y)$?

(a) 0.5 (b) –0.5 (c) 1 (d) 0

13. What is the value of the rank correlation coefficient if $\Sigma d^2 = 0$?

(a) 0 (b) –1 (c) 1 (d) 0.5

14. In the method of rank correlation, in usual notations if $R_x = R_y$ for each pair of observations then what is the vlue of the r ?

(a) 0 (b) –1 (c) 1 (d) 0.1

15. In the method of rank correlation, what is the sum of differences of the ranks of two variables ?

(a) 0 (b) –1 (c) 1 (d) Any real number

16. In the method of rank correlation, if the ranks of two variables are exactly in reverse order then what is the value of r ?

(a) $r = 0$ (b) $r = -1$ (c) $r = 1$ (d) $r = 0.1$

17. In usual notations, which term is added in Σd^2 for each repeated observation in the rank correlation ?

(a) $\frac{m^2-1}{12}$ (b) $\frac{m^3-m}{12}$ (c) $\frac{6m^3-m}{12}$ (d) $n\left(n^2-1\right)$

18. Which kind of correlation will you get between the number of units sold and its revenue at constant price ?

(a) Perfect Positive (b) Partial Positive (c) Perfect Negative (d) Partial Negative

Section B

Answer the following questions in one sentence :

1. Define correlation.

2. Define correlation coefficient.

Identify, whether there is a positive correlation or negative correlation between the following pairs of variables (Question 3 to Question 6).

3. The age of an adult person and life insurance premium at the time of taking an insurance under a plan.

4. The sales and profit of last five years for a mostly accepted product of a company.
5. The rate of inflation and the purchase power of common man of a country when income of the common man is stable.
6. Altitude and amount of Oxygen in air.
7. What can be said about the correlation between the annual import of crude oil and the number of marriages during the same time period ?
8. The correlation coefficient between X and Y is 0.4. What will be the value of correlation coefficient if 5 is added in each observation of X and 10 is substracted from each observation of Y ?
9. What is the main limitation of scatter diagram method ?
10. If the value of $n\left(n^2-1\right)$ is six times the value of Σd^2 then what is the value of r ?
11. What will be the sign of r if the value of the covariance is negative ?

Section C

Answer the following questions :

1. Explain the meaning of positive correlation with an illustration.
2. Explain the meaning of negative correlation with an illustration.
3. Write the assumptions of Karl Pearson's method.
4. Define : Scatter Diagram.
5. What is spurious correlation ?
6. Explain the cause and effect relationship.
7. Explain : Perfect positive correlation
8. Explain : Perfect negative correlation
9. When is it necessary to use rank correlation ?
10. In which situation, the values of Karl Pearson's correlation coefficient and Spearman's rank correlation coefficient are equal ?
11. Find the value of r if $Cov(x, y)=120,\ s_x=12,\ s_y=15$.
12. Find the value of r if $\Sigma(x-\bar{x})(y-\bar{y})=-65,\ s_x=3,\ s_y=4$ and $n=10$.
13. For 10 pairs of observations, $\Sigma d^2=120$. Find the value of the rank correlation coefficient.

Section D

Answer the following questions :

1. Explain scatter diagram method.
2. Write merits and limitations of scatter diagram method.
3. Write the properties of correlation coefficient.
4. Write the merits and limitations of Karl Pearson's method.
5. Interpret $r = 1$, $r = -1$ and $r = 0$.
6. Explain Spearman's Rank correlation method.
7. Write merits and limitations of Spearman's rank correlation method.
8. How would you interpret partial correlation ?
9. State the necessary precautions to be taken while interpreting the value of correlation coefficient.
10. The following data is available for two variables rainfall in mm. (X) and yield of crop Qtl/ Hectare (Y).

    $n = 10$, $\bar{x} = 120$, $\bar{y} = 150$, $s_x = 30$, $s_y = 40$ and $\Sigma xy = 189000$. Find the correlation coefficient.
11. The following information is obtained for 9 pairs of observations.

    $\Sigma x = 51$, $\Sigma y = 72$, $\Sigma x^2 = 315$, $\Sigma y^2 = 582$, $\Sigma xy = 408$. Find the correlation coefficient.
12. The information obtained on the basis of ranks given by two judges to eight contestants of a dance competition is given below.

    $\Sigma\left(R_x - R_y\right)^2 = 126$

    Where R_x and R_y are the ranks given to a contestant by the two judges respectively. Find Spearman's rank correlation coefficient.
13. The ranks given by two experts on the basis of interviews of five candidates for a job are (3, 5), (5, 4), (1, 2), (2, 3) and (4, 1). Find the rank correlation coefficient from this data.

Section E

Solve the following :

1. The following information is obtained to study the relation between the selling price of nose mask and its demand during an epidemic.

Price (₹)	38	45	40	42	35
Demand (units)	103	92	97	98	100

Find the correlation coefficient between the price and demand of mask by Karl Pearson's method.

2. In order to study the relationship between the abilities in the subjects of Human Resource Management and Personality Development for the students of a post graduate level course, a sample of 5 students is taken and the following information is obtained.

Student	1	2	3	4	5
Marks in HRM	45	25	40	20	45
Marks in PD	47	23	17	35	48

Calculate the Karl Pearson's correlation coefficient between the marks of both the subjects.

3. A vendor wants to display lipsticks of different brands according to their popularity. For that, he invites two experts Preyal and Nishi to rank the lipsticks of different brands.

Lipstick	A	B	C	D	E	F	G
Rank by Preyal	5	6	7	1	3	2	4
Rank by Nishi	5	7	6	2	1	4	3

Find the rank correlation coefficient to know the smilarity in the decision of both the experts.

4. A merchant wants to study the relation between prices of tea and coffee in Ahmedabad city. He obtains the following information about prices of tea and coffee of the last six months.

Price per kg for tea (₹)	340	370	450	320	300	360
Price per 100 grams for coffee (₹)	190	215	200	180	163	175

Calculate the rank correlation coefficient between the price of tea and coffee.

5. The demand of an imported fruit in a local market is very uncertain. To know the relation between the price of the fruit and its supply, a vendor collects the information about the average price and supply for last ten months.

Average price per unit (₹)	65	68	43	38	77	48	35	30	25	50
Supply (hundred units)	52	53	42	60	45	41	37	38	25	27

Find the rank correlation between the average price and the supply.

6. To know the relation between the results of the examinations taken in a span of short time, a teacher has conducted two examinations in last two weeks and the ranks obtained by seven students are as follows.

Student	A	B	C	D	E	F	G
Rank in Test 1	5	1	2	3.5	3.5	7	6
Rank in Test 2	7	1	4	6	5	3	2

Find the rank correlation coefficient to know the similarity between the results of two examinations.

Section F

Solve the following :

1. The information of fertilizer used (in tons) and productivity (in tons) of eight districts is given below.

Fertilizer (tons)	15	18	20	25	29	35	40	38
Productivity (tons)	85	93	95	105	115	130	140	145

Calculate the correlation coefficient by Karl Pearson's method.

2. Find the Karl Pearson's correlation coefficient from the following information of the average weekly hours spent on Video games and the grade points obtained in an examination by 6 children of a big city.

Weekly average hours spent for Video games	43	47	45	50	40	51
Grade points obtained in an examination	5.2	4.9	5.0	4.7	5.4	4.3

3. Find Karl Pearson's correlation coefficient between density of population (per square km) and death rate (per thousand) from the following data.

City	A	B	C	D	E	F	G
Density (per sq. km)	750	600	350	500	200	700	850
Death rate (per thousand)	30	20	15	20	10	25	50

4. The following information is obtained to study the relationship between the advertisement cost and the sales of electric fans of the companies manufacturing electric fans. Find the correlation coefficient between advertisement cost and the sales by Karl Pearson's method.

Company	A	B	C	D	E	F
Advertisement Cost (lakh ₹)	140	120	80	100	80	180
Sales of electric fans (crore ₹)	35	45	15	40	20	50

5. A doctor obtains the following information for the weights of seven mothers and their children from a maternity home for his research to know the relation between the weights of mother and weights of their children at the time of birth.

Weight of mother (kg)	59	72	66	64	77	66	60
Weight of child (kg)	2.5	3.4	3.1	2.7	2.8	2.3	3.0

Find rank correlation coefficient between the weights of mother and child.

6. The following data is obtained to know the relation between maximum day temperature and the sale of ice-cream in Ahmedabad city.

Maximum Temperature (Celsius)	35	42	40	39	44	40	45	40
Sale of ice cream (kg)	600	680	750	630	920	750	900	720

Calculate the rank correlation coefficient.

7. An entrance test required to study abroad is conducted online. The marks obtained in Reasoning Ability and English Speaking in this online test (having negative marking system for wrong answer) by 5 students selected in a sample are given below.

Student	A	B	C	D	E
Marks in Reasoning Ability	5	5	5	5	5
Marks in English Speaking	2	–2	–2	0	2

Find the rank correlation coefficient between Reasoning Ability and ability in English Speaking.

8. Six dancers A, B, C, D, E and F in a dance competition were judged by two dance Gurus. The ranks assigned to the dancers are as follows.

Rank	1	2	3	4	5	6
By Guru 1	B	F	A	C	D	E
By Guru 2	F	A	C	B	E	D

Find the rank correlation coefficient between the judgement of the two Gurus.

9. The following data is obtained for two variables, inflation (X) and interest rate (Y).

$n = 50,\ \Sigma x = 500,\ \Sigma y = 300,\ \Sigma x^2 = 5450,\ \Sigma y^2 = 2000,\ \Sigma xy = 3090$

Later on, it was known that one pair of observation (10, 6) was included additionally by mistake. Find the correlation coefficient by excluding this pair of observations.

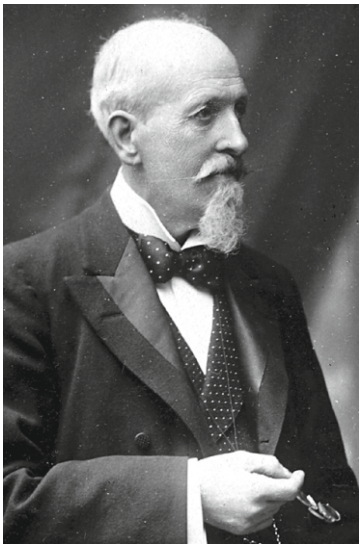
10. The information regarding sales (X) and expenses (Y) of 10 firms is given below.

$\bar{x} = 58,\ \bar{y} = 14,\ \Sigma(x-65)^2 = 850,\ \Sigma(y-13)^2 = 32,\ \Sigma(x-65)(y-13) = 0$

Find the correlation coefficient.

11. Daily calorie intake of ten persons is X and their weight is Y kg. The rank correlation coefficient from this information is 0.6. On subsequent verification, it was noticed that the difference of ranks of X and Y for one of the persons was taken as 2 instead of 4. Find the correct value of rank correlation coefficient.

12. The information of health index x and life expectancy y is obtained for 10 people. These data are ranked to find the rank correlation coefficient and the sum of squares of the ranks was found to be 42.5. It was also observed that health index 70 was repeated three times and life-expectancy 45 was repeated twice in the data. Find the rank correlation coefficient using this information.



Charles Edward Spearman
(1863 –1945)

Charles Edward Spearman was an English psychologist known for work in statistics, as a pioneer of factor analysis and for Spearman's rank correlation coefficient. He also did seminal work on models for human intelligence, including his theory that disparate cognitive test scores reflect a single General intelligence factor and coining the term g-factor.

After serving army for 15 years, he went on to study for a Ph.D. in experimental psychology. Spearman joined University College London and stayed there until he retired in 1931. Initially he was Reader and head of the small psychological laboratory. In 1911 he was promoted to the Grote professorship of the Philosophy of Mind and Logic. His title changed to Professor of Psychology in 1928 when a separate Department of Psychology was created.

His many published papers cover a wide field, but he is especially distinguished by his pioneer work in the application of mathematical methods to the analysis of the human mind and his original studies of correlation in this sphere.

"Prediction is very difficult, especially about the future."

– Niels Bohr

3

Linear Regression

Contents :

3.1 Introduction

In the previous chapter 2, we have studied the concept of correlation. We have seen whether the correlation between two variables is positive or negative is known by the correlation coefficient. Moreover, we get numerical measure of the closeness of the variables. But the coefficient of correlation fails to provide the expected value of one variable for the given value of the other variable. When some relation exists between two variables, many times it is necessary to obtain the approximate or estimated value of one variable for a known value of the other variable using this relation.

e.g. We know that there is a correlation between advertisement cost and sale of an item. Now, for some given amount of advertisement cost, if we want to know the corresponding expected sale then it is not possible to obtain it only by correlation. For this, it is necessary to use the concept of regression.

The literal meaning of regression is 'to avert' or 'return to the mean value'. The term regression was first used by a statistician Sir Francis Galton during his study of human inheritance. He had collected the information about the height of 1000 pairs of fathers and sons. He revealed the following interesting results.

(i) Tall fathers have tall sons and short fathers have short sons.

(ii) The average height of sons of a group of tall fathers is less than average height of group of tall fathers.

(iii) The average height of sons of a group of short fathers is greater than average height of group of short fathers.

So, it is clear from the above findings that the heights of sons show regressive tendency with respect to the height of their fathers. The existence of this tendency restricts the humans to split into two races of pigmies and giants. So, Sir Francis Galton has given the name regression to describe such relation.

Regression is a functional relation between two correlated variables. We shall study the concept of regression under the assumption that there exists a cause-effect relationship between two variables.

3.2 Linear Regression Model

A set of one or more equations representing a relation or a problem is called a model. A statistical model which describes the cause and effect relationship between two variables is called a regression model. Generally, out of two variables having cause-effect relationship, the causal variable is denoted by X. We shall call this variable as independent or explanatory variable and effect variable is denoted by Y. We shall call this variable as dependent or explained variable. Let us understand the meaning of independent variable and dependent variable from the following illustrations :

(i) In case of 'advertisement cost' and 'sales', generally, because of increase (decrease) in the 'advertisement cost', corresponding 'sales' also increases (decreases), so we shall take 'advertisement cost' as independent variable X and 'sales' as dependent variable Y.

(ii) In case of 'rainfall' and 'yield of rice' in some region, it is very clear that 'yield of rice' depends on 'rainfall'. So, we shall take 'rainfall' as independent variable X and 'yield of rice' as dependent variable Y.

In a regression model, the dependent variable Y is expressed in the form of an appropriate mathematical function of the independent variable X.

Now, we shall define a linear regression model as follows.

$$Y = \alpha + \beta X + u$$

Where, Y = Dependent Variable

X = Independent Variable

α = Constant

β = Constant

u = Disturbance Variable of the Model

The inadequacy of the linearity between two variables X and Y is shown by u. The perfect linear relation is possible in natural science like mathematics. So, the disturbance variable u obviously becomes 0 in such a case. In other words, when there is a perfect linear correlation between two variables X and Y then the regression model is $Y = \alpha + \beta X$. But we know that exact linear relation between the variables is not always possible in business, economics and social science as these correlated variables are also affected by other factors. Thus, when there is a partial correlation between the variables X and Y then the linear regression model is $Y = \alpha + \beta X + u$. From the above discussion, we can define linear regression in simple words as follows.

"A mathematical or functional relationship between two correlated variables which helps in estimating the value of dependent variable for some given (known) value of independent variable is called **Linear Regression.**''

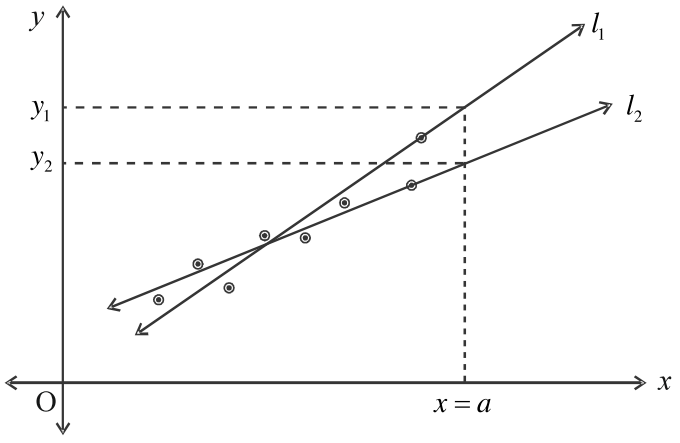
3.3 Fitting of Regression Line

In a scatter diagram of two correlated variables, if the points are clustered around a line, we can say that there is a linear regression. The method of obtaining such a line expressing the relation between two variables is called fitting of a regression line.

There are two methods for fitting a regression line : (1) Method of Scatter Diagram (2) Method of Least Squares.

3.3.1 Method of Scatter Diagram

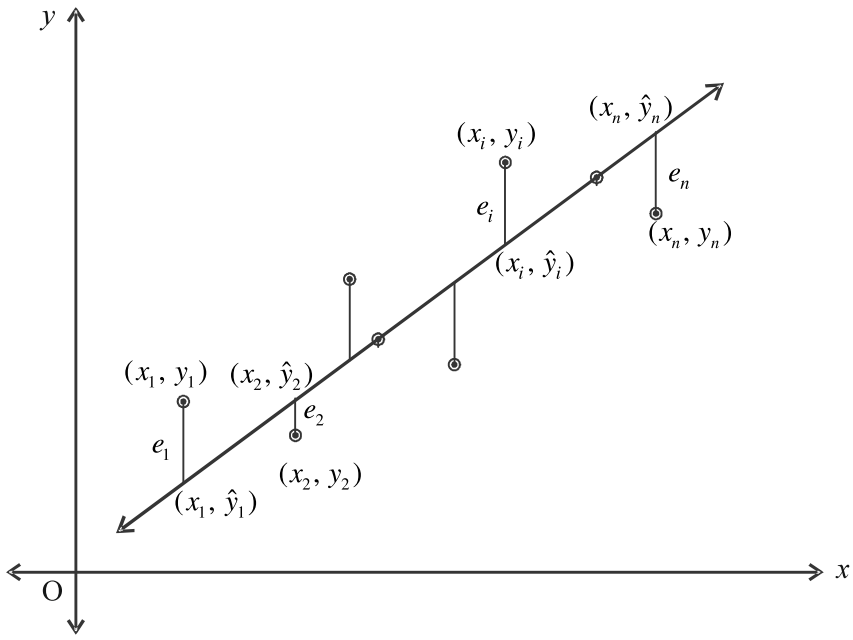
Suppose n ordered pairs of observations of two correlated variables X and Y are $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Using this data, we draw a scatter diagram. Now, a line is drawn in such a way that it is close to almost all the points of the scatter diagram. If Y is a dependent variable and X is an independent variable then such a line is called regression line of Y on X and an approximate value of dependent variable Y can be obtained for any given value of independent variable X from it. Since no computation is required to draw such a line, it is very easy and quick method of fitting a regression line. But there is a problem in doing so. Different persons may draw different lines. As a result, different persons may provide different estimates of the dependent variable Y for the same value of independent variable X. It can be seen very easily from the following scatter diagram.



Two different persons have drawn two different lines l_1 and l_2 in the following scatter diagram of the same data. We can see that for some value 'a' of independent variable X, corresponding estimated value is 'y_1' from line l_1 and it is 'y_2' from the line l_2. Thus we get different estimates for dependent variable Y from different lines for a single value of independent variable X. So, it can be said that this method is subjective. A line of regression drawn by this method is not the best fitted line because it does not guarantee the best estimate of the dependent variable. The method of least square is used to obtain such a best fitted regression line.

3.3.2 Method of Least Squares

Suppose n ordered pairs of observations of two correlated variables X (independent variable) and Y (dependent variable) are $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. We shall draw a scatter diagram for this data to understand the method of least squares.



If an equation of the best fitted line describing the linear regression between the variables X and Y is $\hat{y} = a + bx$ then the constants a and b of this line can be obtained by the method of least squares as follows.

Let the estimated values of variable Y corresponding to values $x_1, x_2, x_3, ..., x_n$ of variable X are $\hat{y}_1, \hat{y}_2, \hat{y}_3, ... \hat{y}_n$ from the line and the corresponding observed values of Y are $y_1, y_2, y_3, ... y_n$ respectively. Now, for some $X = x_i$, estimated value of Y from the line is $\hat{y}_i = a + bx_i$. The vertical distance (i.e. distance parallel to Y-axis) between observed value y_i and the estimated value $\hat{y}_i$ is called an error in the estimation. It is denoted by e_i.

$$\therefore\ e_i = y_i - \hat{y}_i = y_i - (a + bx_i) = y_i - a - bx_i$$

Where, $i = 1, 2, 3, ..., n$

Obviously, the error will be positive for points above the line the error will be negative for points below the line and it will be zero for the points which are on the line.

Now, the values of constants a and b of the fitted line $\hat{y} = a + bx$ (known as regression line of Y on X) are obtained in such a way that the sum of the squares of the errors is minimum.

i.e. $\Sigma e_i^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - a - bx_i)^2$ is minimum.

By ignoring the suffix i for convenience, we can get such values of a and b by a simple algebraic method, which are as follows.

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

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

And

$$a = \bar{y} - b\,\bar{x}$$

The line $\hat{y} = a + bx$ obtained by this method is a line passing as close as possible to the points of scatter diagram. The sum of squares of the errors is minimised while obtaining the regression line. Therefore, this method is called **'method of least squares'.**

The value of b obtained by this method is called the regression coefficient of the regression line of Y on X. b is also called slope of the regression line and the constant a is called intercept of the regression line.

Interpretation of regression coefficient b

b = the estimated change in the value of Y for a unit change in the value of X.

i.e. when $b > 0$, it means that a unit increase in the value of independent variable X implies an estimated increase of b units in the value of dependent variable Y.

when $b < 0$, it means that a unit increase in the value of independent variable X implies an estimated decrease of $|b|$ units in the value of dependent variable Y.

Note that the regression line obtained by the method of least squares is also known as the line of best fit.

Note : (1) The regression coefficient b can also be denoted by b_{yx}. If not required, generally we shall denote regression coefficient by b only.

(2) If all the points in a scatter diagram are on one line only then error will be zero for all the points. Hence, the estimated value $\hat{y}$ is same as its observed value y. So, the form of the regression line will be $y = a + bx$ in place of $\hat{y} = a + bx$. Naturally, in this situation r is 1 if $b > 0$ and r is -1 if $b < 0$.

Additional Information for understanding

Generally, only 'fitted line' is mentioned for the regression line obtained instead of 'best fitted line'.

Now, let us take some examples to obtain a regression line.

Illustration 1 : The following observations are obtained for life (years of usage) of cars and their average annual maintenance costs of a specific model of car of a particular company.

Life of cars (years)	2	4	6	8
Average annual maintenance cost (thousand ₹)	10	20	25	30

Obtain the regression line of maintenance cost on the life of cars. Also, estimate the maintenance cost if the life of a car is 10 years.

'Life of a car' is an independent variable. So, we shall denote it by variable X and 'maintenance cost' is dependent variable. So, we shall denote it by Y. Considering at the data, we shall prepare the following table for obtaining the regression line.

	Life of car (years) x	**Maintenance cost (thousand ₹) y**	xy	x^2
	2	10	20	4
	4	20	80	16
	6	25	150	36
	8	30	240	64
Total	**20**	**85**	**490**	**120**

$\bar{x} = \frac{\Sigma x}{n} = \frac{20}{4} = 5, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{85}{4} = 21.25$

Let us find the regression coefficient as follows.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{4(490) - (20)(85)}{4(120) - (20)^2}$$

$$= \frac{1960 - 1700}{480 - 400}$$

$$= \frac{260}{80}$$

$$= 3.25$$

$\therefore \quad b = 3.25$

Now, putting the values of $\bar{x}, \bar{y}$ and b in the formula of a,

$$a = \bar{y} - b\,\bar{x}$$

$$= 21.25 - 3.25(5)$$

$$= 21.25 - 16.25$$

$\therefore \quad a = 5$

So, the regression line of 'maintenance cost' (Y) on 'life of car' (X) is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = 5 + 3.25\,x$

Putting $X = 10$,

$$\hat{y} = 5 + 3.25(10)$$

$$= 5 + 32.5 = 37.5$$

$\therefore \quad \hat{y} = 37.5$

So, when the life of a car is 10 years then its estimated maintenance cost is ₹ 37.5 thousand **Note :** Since $b = 3.25$, we can say that every year (one unit change in X), the maintenance cost of the car increases by approximately ₹ 3.25 thousand (change in Y).

Illustration 2 : The monthly sale of different types of laptops (in hundred units) and its profit (in lakh ₹) for the last six months for a company is given below.

Month	1	2	3	4	5	6
No. of laptops sold (hundred units) x	5	7	5	12	8	3
Profit (lakh ₹) y	8	9	10	15	10	6

Obtain the regression line of Y on X. Also find the error in estimating Y for X = 7.

	No. of laptops sold (hundred units) x	Profit (lakh ₹) y	xy	x^2
	5	8	40	25
	7	9	63	49
	5	10	50	25
	12	15	180	144
	8	10	80	64
	3	6	18	9
Total	**40**	**58**	**431**	**316**

$\bar{x} = \frac{\Sigma x}{n} = \frac{40}{6} = 6.67;\ \bar{y} = \frac{\Sigma y}{n} = \frac{58}{6} = 9.67$

Let us find the regression coefficient b as follows.

$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{6(431) - (40)(58)}{6(316) - (40)^2}$$

$$= \frac{2586 - 2320}{1896 - 1600}$$

$$= \frac{266}{296}$$

$$= 0.8986$$

$$\simeq 0.90$$

$\therefore b \simeq 0.90$

By putting the values of $\bar{x}, \bar{y}$ and b in the formula of a,

$$a = \bar{y} - b\bar{x}$$

$$= 9.67 - 0.90(6.67)$$

$$= 9.67 - 6.003$$

$$= 3.667$$

$\therefore a \simeq 3.67$

So, regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \hat{y} = 3.67 + 0.9x$

Now, to find the error for $X = 7$, first we obtain the estimated value of Y corresponding to it.

Putting $X = 7$,

$\hat{y} = 3.67 + 0.9(7)$

$= 3.67 + 6.3$

$\therefore \hat{y} = ₹\,9.97$ lakh

Now, we can see from the available data that observed value of Y when $X = 7$ is 9.

$\therefore$ Error $e = y - \hat{y}$

$= 9 - 9.97$

$\therefore e = ₹ - 0.97$ lakh

Illustration 3 : In order to study the relationship between the repairing time of accident damaged cars and the cost of repair, the following information is collected.

Repairing time of a car (man hours)	32	40	25	29	35	43
Repairing cost (thousand ₹)	25	35	18	22	28	46

Obtain the regression line of Y (repairing cost) on X (repairing time). If the time taken to repair a car is 50 hours, find an estimate of the repairing cost.

Here, $n = 6, \bar{x} = \frac{\Sigma x}{n} = \frac{204}{6} = 34$ અને $\bar{y} = \frac{\Sigma y}{n} = \frac{174}{6} = 29$

	Repairing time (man hours) x	**Repairing cost (thousand ₹) y**	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$	$(x - \bar{x})^2$
	32	25	–2	–4	8	4
	40	35	6	6	36	36
	25	18	–9	–11	99	81
	29	22	–5	–7	35	25
	35	28	1	–1	–1	1
	43	46	9	17	153	81
Total	**204**	**174**	**0**	**0**	**330**	**228**

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$= \frac{330}{228}$

$= 1.4474$

$\simeq 1.45$

$\therefore b \simeq 1.45$

Now, by putting the value of $\bar{x}$, $\bar{y}$ and b in the formula of a,

$a = \bar{y} - b\bar{x}$

$= 29 - 1.45(34)$

$= 29 - 49.3$

$\therefore a = -20.3$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = -20.3 + 1.45x$$

Putting $X = 50$,

$$\hat{y} = -20.3 + 1.45(50)$$

$$= -20.3 + 72.5$$

$$\therefore \quad \hat{y} = 52.2$$

So, when the repairing time is 50 hours, the estimated repairing cost is ₹ 52.2 thousand.

Exercise 3.1

1. From the following data of price (in ₹) and demand (in hundred units) of a commodity, obtain the regression line of demand on price. Also estimate the demand when price is 20 ₹.

Price (₹)	12	14	15	16	18	21
Demand (hundred units)	18	12	10	8	7	5

2. To study the relationship between the time of usage of cars and its average annual maintenance cost, the following information is obtained :

Car	1	2	3	4	5	6
Time of usage of a car (years) x	3	1	2	2	5	3
Average annual maintenance cost (thousand ₹) y	10	5	8	7	13	8

Obtain the regression line of Y on X. Find an estimate of average annual maintenance cost when the usage time of a car is 5 years. Also find its error.

3. The information for a year regarding the average rainfall (in cm) and total production of crop (in tons) of five districts is given below :

Average rainfall (cm)	25	32	38	29	31
Crop (tons)	84	90	95	88	93

Find the regression line of production of crop on rainfall and estimate the crop if average rainfall is 35 cm.

4. The following data gives the experience of machine operators and their performance ratings.

Operator	1	2	3	4	5	6	7	8
Experience (years) x	12	5	10	3	18	4	12	16
Performance rating y	83	75	80	78	89	68	88	87

Calculate the regression line of performance ratings on the experience and estimate the performance rating of an operator having 7 years of experience.

*

3.4 Utility of the Study of Regression

The following are some utilities of regression :

(1) We can determine a functional relation between two correlated variables.

(2) Once the functional relation is established, it can be used to predict the unknown value of dependent variable Y on the basis of known value of independent variable X.

(3) We can determine the approximate change in the value of dependent variable Y for a unit change in the value of independent variable X.

(4) We can determine the error in the estimation of dependent variable obtained by a regression line.

Regression is very useful for economists, planners, businessmen, administrators, researchers, etc.

Short-cut Method for computing Regression coefficient

When the values of X and Y are relatively large and / or fractional, it is difficult to calculate the terms like x^2, xy, etc. In such cases, an alternative formula can be used. It is based on the following property of regression coefficient.

Property : The regression coefficient is independent of change of origin but not of change of scale.

If $b = b_{yx}$ is a regression coefficient of a regression line of Y on X then using the above property, the following formulae can be written for the regression coefficient by short-cut method.

(1) If $u = x - A$ and $v = y - B$ then

$$b = b_{yx} = b_{vu} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$$

(2) If $u = \frac{x-A}{c_x}$ and $v = \frac{y-B}{c_y}$ then

$$b = b_{yx} = b_{vu} \cdot \frac{c_y}{c_x} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

Here, A, B, c_x and c_y are constants and $c_x > 0, c_y > 0$.

Illustration 4 : In order to determine the relationship between monthly income (in thousand ₹) and monthly expenditure (in thousand ₹) of people of a group, a sample of seven persons is taken from that group and the following information is obtained.

Person	1	2	3	4	5	6	7
Monthly income (thousand ₹)	60	70	64	68	62	65	72
Monthly expenditure (thousand ₹)	50	59	57	50	53	58	60

Obtain the regression line of monthly expenditure on monthly income. If a person of the group has monthly income of ₹ 75 thousand, estimate his monthly expenditure.

Since the regression line of monthly expenditure on monthly income is to be obtained, we shall take 'monthly expenditure' as variable Y and 'monthly income' as variable X.

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{461}{7} = 65.86$ and $\bar{y} = \frac{\Sigma y}{n} = \frac{387}{7} = 55.29$

So, by taking $A = 65$ and $B = 55$, we can define u and v as follows.

$u = x - A = x - 65$ and $v = y - B = y - 55$

	Monthly income (thousand ₹) x	**Monthly expenditure (thousand ₹) y**	$u = x - 65$	$v = y - 55$	uv	u^2
	60	50	–5	–5	25	25
	70	59	5	4	20	25
	64	57	–1	2	–2	1
	68	50	3	–5	–15	9
	62	53	–3	–2	6	9
	65	58	0	3	0	0
	72	60	7	5	35	49
Total	**461**	**387**	**6**	**2**	**69**	**118**

b can be obtained by short-cut method as follows.

$$b = b_{yx} = b_{vu} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2}$$

$$= \frac{7(69) - (6)(2)}{7(118) - (6)^2}$$

$$= \frac{483 - 12}{826 - 36}$$

$$= \frac{471}{790}$$

$$= 0.5962$$

$$\therefore \quad b \simeq 0.60$$

Now, $a = \bar{y} - b\bar{x}$

$$= 55.29 - 0.60(65.86)$$

$$= 55.29 - 39.516$$

$$= 15.774$$

$$\therefore \quad a = 15.77$$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = 15.77 + 0.60x$$

Putting $X = 75$,

$$\hat{y} = 15.77 + 0.60(75)$$

$$= 15.77 + 45$$

$$= 60.77$$

$$\therefore \quad \hat{y} = 60.77$$

So, if a person has monthly income of ₹ 75 thousand, his approximate monthly expenditure is ₹ 60.77 thousand.

Illustration 5 : For the data given in illustration 1, obtain the regression line of maintenance cost (Y) on the life of cars (X) by using short-cut method.

Life of car (years) x	**2**	**4**	**6**	**8**
Annual maintenance cost (thousand ₹) y	**10**	**20**	**25**	**30**

All the values of X here are divisible by 2 and that of Y are divisible by 5. Moreover $\bar{x} = 5$ and $\bar{y} = 21.25$. So, we shall take $A = 4, B = 20, c_x = 2, c_y = 5.$

Now, let us define u and v as follows :

$$u = \frac{x-A}{c_x} = \frac{x-4}{2} \quad \text{and} \quad v = \frac{y-B}{c_y} = \frac{y-20}{5}$$

	x	y	$u = \frac{x-4}{2}$	$v = \frac{y-20}{5}$	uv	u^2
	2	10	–1	–2	2	1
	4	20	0	0	0	0
	6	25	1	1	1	1
	8	30	2	2	4	4
Total	**20**	**85**	**2**	**1**	**7**	**6**

$$b = b_{vu} \cdot \frac{c_y}{c_x} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

$$= \frac{4(7) - 2(1)}{4(6) - (2)^2} \times \frac{5}{2}$$

$$= \frac{28-2}{24-4} \times \frac{5}{2}$$

$$= \frac{26}{20} \times \frac{5}{2}$$

$$b = 3.25$$

Now, $a = \bar{y} - b\bar{x} = 21.25 - 3.25(5) = 21.25 - 16.25 = 5$

$\therefore$ The regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = 5 + 3.25x$$

Note : We can see that, $b_{vu} = \frac{26}{20} = 1.3$ but when it is multiplied by $\frac{c_y}{c_x} = \frac{5}{2}$ then we get $b = 1.3 \times \frac{5}{2} = 3.25$ (as obtained in illustration 1). So, we can understand that when the scale of variable X and/or Y are changed, it is necessary to multiply b_{vu} by $\frac{c_y}{c_x}$ to obtain b.

Illustration 6 : A sample of seven students is taken from the students coming from abroad in the current year to study in university of the Gujarat State. The information regarding their I.Q. and the marks obtained in an examination of 75 marks is given below.

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**
I.Q. x	85	95	100	90	110	125	70
Marks y	46	50	50	45	60	70	40

Obtain the regression line of Y on X and estimate the marks of a student whose I.Q. is 120. Also find the error in estimation when I.Q. is 100.

Here, $n = 7$, $\bar{x} = \frac{\Sigma x}{n} = \frac{675}{7} = 96.43$, $\bar{y} = \frac{\Sigma y}{n} = \frac{361}{7} = 51.57$

Since the values of X and Y are large, means are fractional and all the values of X are divisible by 5, we shall use short-cut method.

By taking $A = 95, B = 50, c_x = 5, c_y = 1$, we define u and v as follows.

$$u = \frac{x-A}{c_x} = \frac{x-95}{5} \quad \text{and} \quad v = \frac{y-B}{c_y} = \frac{y-50}{1} = y - 50$$

	I.Q. x	**Marks y**	**$u = \frac{x-95}{5}$**	**$v = y - 50$**	**uv**	**u^2**
	85	46	–2	–4	8	4
	95	50	0	0	0	0
	100	50	1	0	0	1
	90	45	–1	–5	5	1
	110	60	3	10	30	9
	125	70	6	20	120	36
	70	40	–5	–10	50	25
Total	**675**	**361**	**2**	**11**	**213**	**76**

$$b = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

$$= \frac{7(213) - (2)(11)}{7(76) - (2)^2} \times \frac{1}{5}$$

$$= \frac{1491 - 22}{532 - 4} \times \frac{1}{5}$$

$$= \frac{1469}{528} \times \frac{1}{5}$$

$$= \frac{1469}{2640}$$

$$= 0.5564$$

$\therefore \quad b \simeq 0.56$

Now, $a = \bar{y} - b\bar{x}$

$$= 51.57 - 0.56(96.43)$$

$$= 51.57 - 54.0008$$

$$= -2.4308$$

$\therefore \quad a \simeq -2.43$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = -2.43 + 0.56x$

Putting $X = 120$,

$$\hat{y} = -2.43 + 0.56(120)$$

$$= -2.43 + 67.2$$

$\therefore \quad \hat{y} = 64.77$ marks.

So, when the I.Q. of a student is 120 then his marks are approximately 65.

Now, to obtain the error when I.Q. (X) = 100, first we have to find the estimate of Y i.e. $\hat{y}$.

$$\hat{y} = -2.43 + 0.56x$$

Taking $X = 100$,

$$\hat{y} = -2.43 + 0.56(100)$$

$$= -2.43 + 56$$

$\therefore \quad \hat{y} = 53.57$ marks

But the observed value of Y for $X = 100$ is 50. (See the given data)

$\therefore$ Error $e = y - \hat{y}$

$= 50 - 53.57$

$\therefore$ $e = -3.57$ marks

Note : It is necessary to keep in mind that the error can be obtained only for those values of independent variable (X), for which the observed value of dependent variable (Y) are known.

In this example, we can not obtain the error in estimating Y for $X = 120$ because the observed value of Y when $X = 120$ is not known.

Illustration 7 : From the data and calculation of illustration 12 of the chapter of linear correlation, obtain the regression line of profit on the sales. Estimate the profit when sales is ₹ 3 crore.

From the illustration, we know that

$u = \frac{x-A}{c_x} = \frac{x-2}{0.1}$ and $v = \frac{y-B}{c_y} = \frac{y-5600}{100}$

$\therefore$ $c_x = 0.1$ and $c_y = 100$

Note that c_x is the divisor of $(x - A)$. So, though we have multiplied $(x - A)$ by 10 for simplicity of calculation, c_x is $\frac{1}{10} = 0.1$.

$(\because$ To multiply by 10 is same as to divide by $\frac{1}{10} = 0.1)$

Now $b = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$

$= \frac{9(121) - (0)(1)}{9(60) - (0)^2} \times \frac{100}{0.1}$

$= \frac{1089}{540} \times \frac{100}{0.1}$

$= \frac{108900}{54}$

$= 2016.6667$

$\therefore$ $b \simeq 2016.67$

Now, $a = \bar{y} - b\bar{x}$

$= 5611.11 - 2016.67(2)$

$= 5611.11 - 4033.34$

$\therefore$ $a = 1577.77$

So, the regression line of profit (Y) on the sales (X) is

$\hat{y} = a + bx$

$\therefore$ $\hat{y} = 1577.77 + 2016.67x$

Putting $X = 3$,

$$\hat{y} = 1577.77 + 2016.67(3)$$

$$= 1577.77 + 6050.01$$

$$\therefore \quad \hat{y} = 7627.78$$

So, when sales is ₹ 3 crore then the estimated profit is 7627.78 (thousand ₹).

Activity

Collect the information of monthly income and monthly expenditure of your family from June to December of a year in which you are studying in standard 12. Obtain the regression line of monthly expenditure on the monthly income. Estimate the monthly expenditure of January of the successive year. Check the actual expenditure at the end of January and find the error in your estimation.

3.5 Regression coefficient from covariance and correlation coefficient

When the summary measures like mean, standard deviation (or variance), covariance, correlation coefficient are known for bivariate data of two variables X and Y, regression coefficient and the line of regression can be obtained as follows.

(1) When the measures like $\bar{x}, \bar{y}, s_x^2$ (or s_x), s_y^2 (or s_y) and Cov(x, y) are known,

$$b = \frac{\text{Covariance}(x, y)}{\text{Variance of } x} = \frac{Cov(x, y)}{s_x^2}$$

and $a = \bar{y} - b\bar{x}$

where, $Cov(x, y) = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \dfrac{\Sigma xy - n\,\bar{x}\,\bar{y}}{n}$

$$s_x^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$$s_y^2 = \frac{\Sigma(y-\bar{y})^2}{n} = \frac{\Sigma y^2}{n} - \left(\frac{\Sigma y}{n}\right)^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$$

(2) When the measures like $\bar{x}, \bar{y}, r, s_x$ (or s_x^2), and s_y (or s_y^2) are known,

$$b = r \cdot \frac{\text{S.D. of } y}{\text{S.D. of } x} = r \cdot \frac{s_y}{s_x}$$

and $a = \bar{y} - b\bar{x}$

The regression line of Y on X i.e. $\hat{y} = a + bx$ can be obtained by putting the values of a and b.

Now, we consider some examples in which some summary measures are known and the regression line is to be obtained.

Illustration 8 : The following measures are obtained to study the relation between rainfall in cm (X) and yield of Bajri in Quintal per Hectare (Y) in ten different regions during monsoon.

$n = 10, \bar{x} = 40, \bar{y} = 175, s_x = 12, Cov(x, y) = 360$

Obtain the regression line of yield Y on rainfall X.

Here, $Cov(x, y) = 360$ and $s_x = 12 \therefore s_x^2 = 144$

$$b = \frac{Cov(x, y)}{s_x^2}$$

$$= \frac{360}{144}$$

$\therefore \quad b = 2.5$

and $a = \bar{y} - b\bar{x}$

$$= 175 - 2.5(40)$$

$$= 175 - 100$$

$\therefore \quad a = 75$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = 75 + 2.5x$

Illustration 9 : To study the relation between two variables, yearly income (X) of a family and their yearly investment (Y) in mutual funds, the following information is shown for a sample of 100 families of a city.

X = Annual income of a family (lakh ₹)

Y = Annual investment in mutual fund of a family (thousand ₹)

$\bar{x} = 5.5, \bar{y} = 40.5, s_x = 1.2, s_y = 12.8, r = 0.65$

Obtain the regression line of annual investment in mutual fund of a family on their annual income. Estimate the annual investment in mutual fund of a family whose annual income is ₹ 4.5 lakh.

Here, $n = 100, \bar{x} = 5.5, \bar{y} = 40.5$

$s_x = 1.2, s_y = 12.8$ and $r = 0.65$

Now, $b = r \cdot \frac{s_y}{s_x}$

$$= 0.65 \times \frac{12.8}{1.2}$$

$$= 6.9333$$

$\therefore \quad b \simeq 6.93$

And $a = \bar{y} - b\bar{x}$

$= 40.5 - 6.93\ (5.5)$

$= 40.5 - 38.115$

$= 2.385$

$\therefore \quad a \simeq 2.39$

So, the regression line of Y on X is

$\hat{y} = a + bx$

$\therefore \quad \hat{y} = 2.39 + 6.93x$

Putting $X = 4.5$,

$\hat{y} = 2.39 + 6.93(4.5)$

$= 2.39 + 31.185$

$= 33.575$

$\therefore \quad \hat{y} \simeq 33.58$

So, when annual income of a family is ₹ 4.5 lakh then estimated investment in mutual fund is ₹ 33.58 thousand.

Illustration 10 : The information of price (in ₹) of a ballpen and the supply of ballpen (in units) at the end of each month of a year for a company making ball pen is given below. Estimate the supply of ballpen when its price is ₹ 40.

Detail	**Price (x)**	**Supply (y)**
Average	30	500
Variance	25	10,000
$r = 0.8$		

Here, $\bar{x} = 30, \bar{y} = 500, s_x^2 = 25, s_y^2 = 10000$ and $r = 0.8$

Since $s_x^2 = 25$, $s_x = 5$

Since $s_y^2 = 10000$, $s_y = 100$

Since the supply Y is to be estimated for the price $X = 40$, we shall obtain the regression line of Y on X.

$b = r \cdot \frac{s_y}{s_x}$

$= 0.8 \times \frac{100}{5}$

$\therefore \quad b = 16$

$a = \bar{y} - b\bar{x}$

$= 500 - 16\ (30)$

$= 500 - 480$

$\therefore \quad a = 20$

So, the regression line of Y on X is

$\hat{y} = a + bx$

$\therefore \quad \hat{y} = 20 + 16x$

Putting $X = 40$,

$\hat{y} = 20 + 16(40)$

$= 20 + 640$

$\therefore \quad \hat{y} = 660$ units

So, when the price is ₹ 40, the estimate of supply is 660 units.

Illustration 11 : A person in a state of South India produces spoons from eatable materials. It can be eaten after using it. He launched such spoons for the purpose of selling in a state on an experimental level. The following results are obtained for the average price (in ₹) and its demand (in hundred units) for the last six months.

$\boldsymbol{n = 6, \Sigma x = 45, \Sigma y = 122, \Sigma x^2 = 439, \Sigma xy = 605}$

Obtain the regression line of the demand (Y) of spoons on the price (X) and estimate the demand of spoons when the price of a spoon is ₹ 10.

Here, $\bar{x} = \frac{\Sigma x}{n} = \frac{45}{6} = 7.5$, $\bar{y} = \frac{\Sigma y}{n} = \frac{122}{6} = 20.33$

$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$

$= \frac{6(605) - (45)(122)}{6(439) - (45)^2}$

$= \frac{3630 - 5490}{2634 - 2025}$

$= \frac{-1860}{609}$

$= -3.0542$

$\therefore \quad b \simeq -3.05$

$a = \bar{y} - b\bar{x}$

$= 20.33 - (-3.05)\ (7.5)$

$= 20.33 + 22.875$

$= 43.205$

$\therefore \quad a \simeq 43.21$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \quad \hat{y} = 43.21 - 3.05x$$

Putting X = 10,

$$\hat{y} = 43.21 - 3.05(10)$$

$$= 43.21 - 30.5$$

$$\therefore \quad \hat{y} = 12.71$$

So, when the price is ₹ 10, estimated demand is 12.71 (hundred units).

Illustration 12 : The electricity is generated by windmill manufactured by a company. The following information is obtained by recording five observations regarding the velocity of wind (km per hour) and generation of electricity (in Watts) by a unit of the company.

Velocity of Wind = X km per hour

Electricity Generation = Y Watts

$\bar{x} = 20,\ \bar{y} = 186,\ \Sigma xy = 23200,\ s_x^2 = 50$

Obtain the regression line of electricity generation (Y) on velocity of wind (X). Estimate the electricity generation if the velocity of wind is 25 km per hour.

Here, $n = 5,\ \Sigma xy = 23200,\ \bar{x} = 20,\ \bar{y} = 186$ and $s_x^2 = 50$

Now, $b = \dfrac{Cov(x, y)}{s_x^2}$

$$= \frac{\Sigma xy - n\,\bar{x}\,\bar{y}}{n \cdot s_x^2}$$

$$= \frac{23200 - 5(20)(186)}{5(50)}$$

$$= \frac{23200 - 18600}{250}$$

$$= \frac{4600}{250}$$

$$\therefore \quad b = 18.4$$

$$a = \bar{y} - b\bar{x}$$

$$= 186 - 18.4\ (20)$$

$$= 186 - 368$$

$$\therefore \quad a = -182$$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore \quad \hat{y} = -182 + 18.4x$

Putting $X = 25$,

$$\hat{y} = -182 + 18.4(25)$$

$$= -182 + 460$$

$\therefore \quad \hat{y} = 278$

So, when the velocity of wind is 25 km per hour, approximately 278 watts electricity is generated.

Exercise 3.2

1. The following information is obtained from a study to know the effect of use of fertilizer on the yield of cotton.

Consumption of fertilizer (10 kg) x	28	35	25	24	20	25	20
Yield of cotton per hectare (Quintals) y	128	140	115	120	105	122	100

Obtain the regression line of Y on X and estimate the yield of cotton per hectare if 300 kg fertilizer is used.

2. To know the relationship between the heights of father and sons, obtain the regression line of height of son on the height of father from the following information of eight pairs of fathers and adult sons.

Height of father (cm) x	167	169	171	168	173	166	167	165
Height of son (cm) y	158	170	169	172	170	168	164	167

Estimate the height of a son whose father's height is 170 cm.

3. From the following information of altitude and the amount of effective Oxygen in air at the place, obtain the regression line of amount of effective Oxygen (Y) on the altitude (X). (305 meter $\simeq$ 1000 feet)

Altitude (305 meter) x	0	1	2	3	4	5	6
Effective Oxygen (%) y	20.9	20.1	19.4	17.9	17.9	17.3	16.6

If the altitude of a place is 7 units (1 unit = 305 meter), estimate the percentage of effective Oxygen in air at that place.

4. The following information is obtained to study the relation between the carpet area in a house and its monthly rent in a city.

Carpet area (square meter) x	55	60	75	80	100	120	140
Monthly rent (₹) y	18,000	19,000	20,000	20,000	25,000	30,000	50,000

Obtain the regression line of Y on X. Estimate the monthly rent of a house having carpet area of 110 square meter.

5. The following sample data is obtained to study the relation between the number of customers visiting a mall per day and the sales (ten thousand ₹).

No. of customers x	50	70	100	70	150	120
Sales (ten thousand ₹) y	2.0	2.0	2.5	1.4	4.0	2.5

Obtain the regression line of Y on X. Estimate the sales of a mall if 80 customers have visited the mall on a particular day.

6. The following information is given for ten firms running business of clothes in a city regarding their average annual profit (in lakh ₹) and average annual administrative cost (in lakh ₹).

Particulars	**Profit (in lakh ₹)** x	**Administrative Cost (in lakh ₹)** y
Mean	60	25
Standard Deviation	6	3
Covariance = 10.4		

Obtain the regression line of Y on X.

7. The following information is obtained to study the relationship between average rainfall (in cm) and the yield of maize (in quintal per hectare) in different talukaa of Gujarat.

Particulars	**Rainfall (cm)** x	**Yield of Maize (Quintal per Hectare)** y
Mean	82	180
Variance	64	225
Correlation coefficient = 0.82		

Estimate the yield of maize when the rainfall is 60 cm.

8. The following results are obtained to study the relation between the price of battery (cell) of wrist watch in rupees (X) and its supply in hundred units (Y).

$n = 10,\ \Sigma x = 130, \Sigma y = 220,\ \Sigma x^2 = 2288,\ \Sigma xy = 3467$

Obtain the regression line of Y on X and estimate the supply when price is ₹ 16.

9. The information regarding maximum temperature (X) and sale of ice-cream (Y) of six different days in summer for a city is given below.

Maximum Temperature = X (in Celsius)

Sale of Ice-cream = Y (in lakh ₹)

$\bar{x} = 40,\ \bar{y} = 1.2,\ \Sigma xy = 306,\ s_x^2 = 20$

Obtain the regression line of sale of ice-cream on maximum temperature. Estimate the sale of ice-cream if the maximum temperature on a day is 42 Celsius.

*

3.6 Coefficient of Determination

We know that regression is a functional relation between two correlated variables and it is useful to estimate the value of dependent variable for some given value of independent variable. The coefficient of determination is a measure to find the reliability of such an estimate.

Suppose the regression line of Y on X is $\hat{y} = a + bx$, then the square of the correlation coefficient between observed values of dependent variable y obtained from the observations and its estimated values $\hat{y}$ which are obtained from the regression line is called the **coefficient of determination.** It is denoted by R^2.

$$\therefore R^2 = \left[r\left(y, \hat{y}\right)\right]^2$$

It can be easily checked that R^2 is same as $r^2(x, y)$ or r^2.

$$R^2 = \left[r\left(y, \hat{y}\right)\right]^2$$

$$= \left[r\left(y, a + bx\right)\right]^2$$

$$= \left[r\left(y, x\right)\right]^2$$

$$= \left[r\left(x, y\right)\right]^2$$

$$\therefore R^2 = r^2$$

$\because$ r is independent of change of origin and scale, so from variable $\hat{y}(= a + bx)$, substracting a and then dividing by b, the value of r will not change.

Since $R^2 = r^2$, we can say that the reliability of an estimate of dependent variable Y largely depends on the correlation coefficient r between two variables X and Y.

If $r = \pm 1$ then $R^2 = r^2 = 1$ and there is a perfect linear correlation between X and Y. So, we can say that the estimates of Y obtained from the regression line are 100 % reliable. But if $r = 0$ then $R^2 = r^2 = 0$ and there is no linear correlation between X and Y. So, we can say that the estimates of Y obtained from the regression line are not reliable.

It is clear from the above discussion that high value of R^2 shows that a good linear correlation exists between two variables X and Y. So, we can check whether the linearity assumption of regression is valid or not from the measure of coefficient of determination (R^2). If the value of R^2 is nearer to 1, the assumption of linearity of regression is valid. But if the value of R^2 is nearer to 0, the assumption of linearity of regression between X and Y is not valid.

How much variation in the dependent variable Y can be explained by the regression line, can be obtained from the coefficient of determination. e.g., If $r = 0.9$ for some data, then coefficient of determination = $(0.9)^2 = 0.81$ and therefore $r^2 \times 100$ % = 81 %. So, it can be said that out of total variation in variable Y, the explanation of 81 % variation is obtained from the regression line. So, we can say that the regression model used for the given data is suitable.

Illustration 13 : The following table shows the experience of technicians (in years) employed at various companies and their monthly salary (in thousand ₹).

Experience (years) x	**12**	**8**	**16**	**20**	**5**	**14**	**10**
Monthly Salary (thousand ₹) y	**22**	**15**	**25**	**30**	**12**	**24**	**20**

Calculate the coefficient of determination and check the validity of the linearity assumption of regression between the years of experience and the monthly salary.

Here, $n = 7,\ \bar{x} = \frac{\Sigma x}{n} = \frac{85}{7} = 12.14,\ \bar{y} = \frac{\Sigma y}{n} = \frac{148}{7} = 21.14$

	Experience (year) x	Monthly salary (thousand ₹) y	xy	x^2	y^2
	12	22	264	144	484
	8	15	120	64	225
	16	25	400	256	625
	20	30	600	400	900
	5	12	60	25	144
	14	24	336	196	576
	10	20	200	100	400
Total	**85**	**148**	**1980**	**1185**	**3354**

Now, $R^2 = r^2 = \left[\frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{n\Sigma y^2 - (\Sigma y)^2}}\right]^2$

$= \left[\frac{7(1980) - (85)(148)}{\sqrt{7(1185) - (85)^2} \cdot \sqrt{7(3354) - (148)^2}}\right]^2$

$= \frac{[13860 - 12580]^2}{[8295 - 7225] \cdot [23478 - 21904]}$

$= \frac{(1280)^2}{(1070) \cdot (1574)}$

$= \frac{1638400}{1684180}$

$= 0.9728$

$\therefore\ R^2 \simeq 0.97$

The value of R^2 is very near to 1. So, we can say that the linearity assumption of regression between the years of experience and the monthly salary is valid.

Note : For the above example, we can also compute R^2 by taking $u = x - A$ and $v = y - B$. Here, A and B are suitable constants.

Illustration 14 : In order to study the relationship between the density of population and the number of persons suffering from skin diseases, the following information is obtained for six cities regarding their density of population (per sq. km) and persons suffering from skin diseases (per thousand).

Density (per sq. km) x	**12,000**	**14,500**	**19,000**	**17,500**	**13,500**	**16,000**
Number of patients (per thousand) y	**80**	**60**	**90**	**80**	**40**	**30**

Obtain the regression line of Y on X. Estimate the number of patients suffering from skin diseases if density of population of a city is 15000 (per sq.km). Examine the reliability of this regression model.

Here, $n = 6, \bar{x} = \frac{\Sigma x}{n} = \frac{92500}{6} = 15416.67$; $\bar{y} = \frac{\Sigma y}{n} = \frac{380}{6} = 63.33$

We can see that the values of variable X are multiple of 500 and that of variable Y are of 10. So, by taking A = 15000, B = 60, c_x = 500, c_y = 10, we shall use short-cut method. Let us define u and v as follows.

$$u = \frac{x-A}{c_x} = \frac{x-15000}{500} \quad \text{and} \quad v = \frac{y-B}{c_y} = \frac{y-60}{10}$$

Density (per sq. km) x	**Number of patients (per thousand) y**	**$u = \frac{x-15000}{500}$**	**$v = \frac{y-60}{10}$**	**uv**	**u^2**	**v^2**
12000	80	–6	2	–12	36	4
14500	60	–1	0	0	1	0
19000	90	8	3	24	64	9
17500	80	5	2	10	25	4
13500	40	–3	–2	6	9	4
16000	30	2	–3	–6	4	9
Total	**92500**	**380**	**5**	**2**	**22**	**139**	**30**

$$b = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{n\Sigma u^2 - (\Sigma u)^2} \times \frac{c_y}{c_x}$$

$$= \frac{6(22) - (5)(2)}{6(139) - (5)^2} \times \frac{10}{500}$$

$$= \frac{132 - 10}{834 - 25} \times \frac{1}{50}$$

$$= \frac{122}{809} \times \frac{1}{50}$$

$$= \frac{122}{40450}$$

$\therefore b \simeq 0.003$

$$a = \bar{y} - b\bar{x}$$
$$= 63.33 - 0.003\,(15416.67)$$
$$= 63.33 - 46.25$$

$\therefore\ a = 17.08$

$\therefore$ The regression line of Y on X is

$$\hat{y} = a + bx$$

$\therefore\ \hat{y} = 17.08 + 0.003\,x$

Putting $X = 15000$,

$$\hat{y} = 17.08 + 0.003(15000)$$
$$= 17.08 + 45$$

$\therefore\ \hat{y} = 62.08$

So, when the density of a city is 15,000 then approximately $62.08 \simeq 62$ patients are suffering from skin diseases.

Now, reliability of a regression model can be examined by coefficient of determination R^2. So, we obtain it.

$$R^2 = r^2 = \left[\frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\cdot\sqrt{n\Sigma v^2 - (\Sigma v)^2}}\right]^2$$

$$= \frac{[6(22) - (5)(2)]^2}{\left[6(139) - (5)^2\right]\left[6(30) - (2)^2\right]}$$

$$= \frac{(122)^2}{(809)(176)}$$

$$= \frac{14884}{142384}$$

$$= 0.1045$$

$\therefore\ R^2 \simeq 0.10$

As the value of R^2 is very near to 0, it can not be said that the regression model is reliable.

3.7 Properties of Regression Coefficient

(1) Correlation coefficient r and regression coefficient b are either both positive or both negative. ($\because$ We know that standard deviations s_x and s_y are always non-negative and $-1 \le r \le 1$. So, from $b = r \cdot \frac{s_y}{s_x}$ it can be understood that the sign of b will be same as that of r.)

(2) Regression coefficient is independent of change of origin but not independent of change of scale. (This property is discussed in detail in the explanation of the short-cut method of calculation of regression coefficient.)

Note : The regression line of Y on X always passes through the point $(\bar{x}, \bar{y})$.

Illustration 15 : Six pairs of father-son are selected in a sample of an experiment to know the relation between the heights of fathers in cm (X) and the heights of their adult sons in cm (Y).

The following results are obtained from it.

$$\Sigma x = 1020,\ \Sigma y = 990,\ \Sigma(x-170)^2 = 60,\ \Sigma(y-165)^2 = 105$$

$$\Sigma(x-170)(y-165) = 45$$

Obtain the regression line of the heights of sons (Y) on the heights of fathers (X). Also verify the reliability of the regression model.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{1020}{6} = 170$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{990}{6} = 165$$

$$\therefore \Sigma(x-170)^2 = \Sigma(x-\bar{x})^2 = 60$$

$$\Sigma(y-165)^2 = \Sigma(y-\bar{y})^2 = 105$$

$$\Sigma(x-170)(y-165) = \Sigma(x-\bar{x})(y-\bar{y}) = 45$$

$$\therefore b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$$= \frac{45}{60}$$

$$\therefore b = 0.75$$

$$a = \bar{y} - b\bar{x}$$

$$= 165 - 0.75(170)$$

$$= 165 - 127.5$$

$$\therefore a = 37.5$$

So, the regression line of Y on X is

$$\hat{y} = a + bx$$

$$\therefore \hat{y} = 37.5 + 0.75x$$

Now, to verify reliability of the regression model, let us obtain the coefficient of determination R^2.

$$R^2 = \left[\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\cdot\sqrt{\Sigma(y-\bar{y})^2}}\right]^2$$

$$= \frac{(45)^2}{(60)(105)}$$

$$= \frac{2025}{6300}$$

$$= 0.3214$$

$$\therefore R^2 \simeq 0.32$$

Since the value of R^2 is nearer to 0, it can not be said that the regression model is reliable.

Illustration 16 : (i) If the regression line of Y on X is $\hat{y} = 12 - 1.5x$ and the mean of X is 6, find the mean of Y. (ii) If the regression line of Y on X is $\hat{y} = 11.5 + 0.65x$ and $\bar{y} = 18$, find $\bar{x}$.

(i) We know that the regression line always passes through a point $(\bar{x}, \bar{y})$. So, the $\hat{y}$ obtained by putting $\bar{x}$ in place of x in the regression line is $\bar{y}$ or the x obtained by putting $\bar{y}$ in place of $\hat{y}$ is $\bar{x}$.

Putting $\bar{x} = 6$ in place of x in $\hat{y} = 12 - 1.5x$,

$\hat{y} = 12 - 1.5(6)$

$\therefore \hat{y} = 12 - 9$

$\therefore \hat{y} = 3$, so $\bar{y} = 3$

Therefore, the mean of Y is 3.

(ii) As per the above discussion, the value of x obtained by putting $\bar{y} = 18$ in place of $\hat{y}$ in $\hat{y} = 11.5 + 0.65x$, we get $\bar{x}$.

By putting $\hat{y} = \bar{y} = 18$ in $\hat{y} = 11.5 + 0.65x$,

$18 = 11.5 + 0.65x$

$\therefore 6.5 = 0.65x$

$\therefore x = \frac{6.5}{0.65}$

$\therefore x = 10$ So $\bar{x} = 10$

Therefore, the mean of X is 10.

Illustration 17 : (i) If $\bar{x} = 5, \bar{y} = 11$ and $b = 1.2$, obtain the regression line of Y on X. (ii) If $\bar{x} = 60, \bar{y} = 75$ and $s_x^2 : Cov(x, y) = 5:3$, obtain the regression line of Y on X and estimate y for $X = 65$ from it.

(i) Here, $b = 1.2$, $\bar{x} = 5$ and $\bar{y} = 11$.

Now, $a = \bar{y} - b\bar{x}$

$\therefore a = 11 - 1.2(5)$

$= 11 - 6$

$\therefore a = 5$

We get the regression line of Y on X as follows.

$\hat{y} = a + bx$

$\therefore \hat{y} = 5 + 1.2x$

(ii) Here, $\bar{x} = 60,\ \bar{y} = 75$ and $s_x^2 : Cov(x, y) = 5:3$

$s_x^2 : Cov(x, y) = 5:3$

$\therefore \quad \frac{s_x^2}{Cov(x, y)} = \frac{5}{3}$ hence, $\frac{Cov(x, y)}{s_x^2} = \frac{3}{5}$

Now, $b = \frac{Cov(x, y)}{s_x^2} = \frac{3}{5} = 0.6$

and $a = \bar{y} - b\bar{x}$

$= 75 - 0.6\,(60)$

$= 75 - 36$

$\therefore \ a = 39$

We get the regression line of Y on X as follows.

$\hat{y} = a + bx$

$\therefore \quad \hat{y} = 39 + 0.6\,x$

Putting $X = 65$,

$\hat{y} = 39 + 0.6\,x$

$= 39 + 0.6\,(65)$

$= 39 + 39$

$\therefore \ \hat{y} = 78$

So, for $X = 65$, the estimated value of Y is 78.

Illustration 18 : The fitted regression line of Y on X is $\hat{y} = 50 + 3.5x$. If an observation $(16, 108)$ is used in fitting of the line, find the error in estimating Y for $X = 16$. (ii) If one observation $(10, 30)$ is used in the fitting of the line $\hat{y} = 22 + 0.8x$, find the error in estimating Y for $X = 10$. What can you deduce from the value of the error ?

(i) Putting $x = 16$ in $\hat{y} = 50 + 3.5x$,

$\hat{y} = 50 + 3.5(16)$

$\therefore \quad = 50 + 56$

$\therefore \ \hat{y} = 106$

And for $X = 16$, corresponding $Y = 108$ is observed.

$\therefore$ Error $e = y - \hat{y}$

$= 108 - 106$

$\therefore \quad e = 2$

So, the error in estimating Y for $X = 16$ is 2.

(ii) Putting $X = 10$ in $\hat{y} = 22 + 0.8x$,

$$\hat{y} = 22 + 0.8(10)$$

$$= 22 + 8$$

$$\therefore \hat{y} = 30$$

And for $X = 10$, corresponding $Y = 30$ is observed.

$\therefore$ Error $e = y - \hat{y}$

$$= 30 - 30$$

$$\therefore e = 0$$

So, the error in estimating Y for $X = 10$ is 0.

Since value of the error is 0, we can say that the point $(10, 30)$ lies on the fitted line $\hat{y} = 22 + 0.8x$.

Note : For a regression line obtained by the method of least squares, the error is positive for the points above the line, negative for the points below the line and it is zero for the points which are on the line.

Illustration 19 : (i) If the regression line of Y on X is $\hat{y} = 25 + 3x$ and $Cov(x, y) = 48$, find the standard deviation of X. Also find coefficient of determination if the standard deviation of Y is 15. (ii) For the regression line given in the above question, how many units should be increased in the value of X to increase approximately 15 units in Y ?

(i) By comparing the regression line of y on x, $\hat{y} = 25 + 3x$ with its general form $\hat{y} = a + bx$, we get regression coefficient $b = 3$. Since $Cov(x, y) = 48$ is given,

$$b = \frac{Cov(x, y)}{s_x^2}$$

$$\therefore \quad 3 = \frac{48}{s_x^2}$$

$$\therefore \quad s_x^2 = 16$$

$$\therefore \quad s_x = 4$$

So, the standard deviation of X is 4.

Now, the standard deviation of Y, $s_y = 15$ is given.

So, coefficient of determination $R^2 = \left[\frac{Cov(x, y)}{s_x \cdot s_y}\right]^2$

$$\therefore \quad R^2 = \left[\frac{48}{4 \times 15}\right]^2 = (0.8)^2 = 0.64$$

Second Method :

$$b = r \cdot \frac{s_y}{s_x}$$

$$\therefore 3 = r \cdot \frac{15}{4}$$

$$\therefore r = \frac{3 \times 4}{15}$$

$$\therefore r = 0.8$$

$$\therefore R^2 = r^2 = (0.8)^2 = 0.64$$

(ii) $\hat{y} = 25 + 3x$ and regression coefficient $b = 3$. It indicates that if the value of X is increased by one unit then estimated value of Y is increased by 3 units. So, if the value of Y is to be increased approximately by 15 units then the value of X should be increased by $\frac{15}{3} = 5$ units.

Illustration 20 : (i) If the regression line is $\hat{y} = \frac{x}{2} + 5$ and $s_y : s_x = 5:8$, find the coefficient of determination. (ii) If the regression line of Y on X is $4x + 5y - 65 = 0$, find the value of regression coefficient b.

(i) By comparing the regression line $\hat{y} = \frac{x}{2} + 5 = \frac{1}{2} \cdot x + 5$ with its general form $\hat{y} = a + bx$, we get $b = \frac{1}{2}$.

Now, $s_y : s_x = 5:8$

$$\therefore \frac{s_y}{s_x} = \frac{5}{8}$$

and $b = r \cdot \frac{s_y}{s_x}$

$$\therefore \frac{1}{2} = r \cdot \frac{5}{8}$$

$$\therefore r = \frac{1}{2} \times \frac{8}{5}$$

$$\therefore r = 0.8$$

$\therefore$ The coefficient of determination $R^2 = r^2 = (0.8)^2 = 0.64$.

(ii) The regression line of Y on X, $4x + 5y - 65 = 0$ is given.

Now, we convert it into its general form.

$$4x + 5y - 65 = 0$$

$$\therefore 5y = 65 - 4x$$

$$\therefore y = \frac{65 - 4x}{5}$$

$$\therefore y = \frac{65}{5} - \frac{4x}{5}$$

$$\therefore y = 13 - 0.8x$$

By comparing it with $\hat{y} = a + bx$, we get $b = -0.8$.