1

Correlation

INTRODUCTION

In our day-to-day life, we find many examples when a mutual relationship exists between two variables, *i.e.*, with fall or rise in the value of one variable, the fall or rise may take place in the value of other variable. For example, price of a commodity rises as the demand for the commodity goes up. Upto a certain time-period, weight of a person increases with the increase in age. Similarly, the temperature rises with the rise in the sun light. These facts indicate that there is certainly some mutual relationship that exists between the demand for a commodity and its price, the age of a person and his weight, and the sunlight and temperature. The correlation refers to the statistical technique used in measuring the closeness of the relationship between the variables.

DEFINITION OF CORRELATION

Some important definitions of correlation are given below:

1. Correlation analysis deals with the association between two or more variables.
—**Simpson and Kafka**
2. If two or more quantities vary in sympathy, so that movement in one tend to be accompanied by corresponding movements in the other, then they are said to be correlated.
—**Conner**
3. Correlation analysis attempts to determine the degree of relationship between variables.
—**Ya-Lun Chou**

Thus, correlation is a statistical technique which helps in analysing the relationship between two or more variables.

UTILITY OF CORRELATION

The study of correlation is of immense significance in statistical analysis and practical life, which is clear from the following points:

(1) Most of variables show same kind of relationship. For example, there is relationship between price and supply, income and expenditure, etc. With the help of correlation analysis, we can measure the degree of relationship in one figure between different variables like supply and price, income and expenditure, etc.

(2) Once we come to know that the two variables are mutually related, then we can estimate the value of one variable on the basis of the value of another. This function is performed by regression technique, which is based on correlation. In other words, the concept of regression is based on correlation.

(3) Correlation is also useful for economists. An economist specifies the relationship between different variables like demand and supply, money supply and price level by way of correlation.

(4) In business, a trader makes the estimation of costs, sales, prices, etc., with the help of correlation and makes appropriate plans.

Thus, in every field of practical life, correlation analysis is extremely useful in making comparative study of two or more related phenomena and analyzing their mutual relationship.

◘ TYPES OF CORRELATION

Main types of correlation are given below:

(1) Positive and Negative Correlation: On the basis of direction of change of the variables correlation can be classified into two types:

(i) Positive Correlation: If two variables X and Y move in the same direction, *i.e.*, if one rises, other rises too and vice versa, then it is a called as positive correlation. Examples of positive correlation are the relationship between price and supply, between money supply and prices, etc.

(ii) Negative Correlation: If two variables X and Y move in opposite direction, *i.e.*, if one rises, other falls, and if one falls, other rises, then it is called as negative correlation. Examples of negative correlation are the relationship between demand and price, investment and rate of interest, etc.

(2) Linear and Curvi-Linear Correlation: On the basis of change in proportion, correlation is of two types:

(i) Linear Correlation: If the ratio of change of two variables X and Y ($\Delta Y / \Delta X$) remains constant throughout, then they are said to be linearly correlated, like as when everytime price of a commodity rises by 20% as often as its price rises by 10%, then such two variables have linear relationship. If values of these two variables are plotted on a graph then all the points will lie on a straight line.

(ii) Curvi-Linear Correlation: If the ratio of change between the two variables is not constant but changing, correlation is said to be curvi-linear, like as when everytime price of a commodity rises by 10%, then sometimes its supply rises by 20%, sometimes by 10% and sometimes by 40%, then non-linear or curvi-linear correlation exists between them. In case of curvi-linear correlation, values of the variables plotted on a graph will give a curve.

(3) Simple Partial and Multiple Correlation: On the basis of number of variables studied correlation may be classified into three types:

(i) Simple Correlation: When we study the relationship between two variables only, then it is called simple correlation. Relationship between price and demand, height and weight, income and consumption, etc., are all examples of simple correlation.

(ii) Partial Correlation: When three or more variables are taken but relationship between any two of the variables is studied, assuming other variables as constant, then it is called partial correlation. Suppose, under constant temperature, we study the relationship between amount of rainfall and wheat yield, then this will be called as partial correlation.

(iii) Multiple Correlation: When we study the relationship among three or more variables, then it is called multiple correlation. For example, if we study the relationship between rainfall, temperature and yield of wheat, then it is called as multiple correlation.

◘ CORRELATION AND CAUSATION

Correlation is a numerical measure of direction and magnitude of the mutual relationship between the values of two or more variables. But the presence of correlation should not be taken as the belief that the two correlated variables necessarily have causal relationship as well. Correlation does not always arise from causal relationship but with the presence of causal relationship, correlation is certain to exist. Presence of high degree of correlation between different variables may be due to the following reasons:

(1) Mutual Dependence: The study of economic theory shows that it is not necessary that only one variable may affect other variable. It is possible that the two variables may affect each other mutually. In such situation, it is difficult to know which one is the cause and which one is the effect. For example, price of a commodity is affected by the forces of demand and supply. According to the law of demand, with the rise in price (other things remaining constant), demand for the commodity will fall. Here rise in price is the cause and fall in demand is the effect. On the other hand, with fall in demand, price of the commodity falls. Here fall in demand is the cause and fall in price is the effect. Thus there may be high degree of correlation between two variables due to mutual dependence, but it is difficult to know which one is the cause and which one is the effect.

(2) Due to Pure Chance: In a small sample it is possible that two variables are highly correlated but in universe, these variables are unlikely to be correlated, such correlation may be due to either the fluctuations of pure random sampling or due to the bias of investigator in selecting the sample. The following example makes the point clear:

Income (in Rs.)	5,000	6,000	7,000	8,000	9,000
Weight (in Kg.)	100	120	140	160	180

In the data as stated above, there is perfect positive correlation between income and weight, *i.e.*, weight increases with rise in income and the rate of change of the two variables is also the same. Still such kind of correlation cannot be said to be meaningful. Such relationship is said to be *spurious or non-sense*.

(3) Correlation Due to any Third Common Factor: Two variables may be correlated due to some common third factor rather than having direct correlation. For example, if there is high degree of positive correlation between per hectare field of tea and rice, then this does not imply that rice yield has risen due to the rich yield of tea. Another reason of the good yield of these two is the good rainfall well in time that affects both of these two.

◘ DEGREE OF CORRELATION

Degree of correlation can be known by coefficient of correlation (r). The following can be various types of the degree of correlation:

(1) Perfect Correlation
(2) High Degree of Correlation
(3) Moderate Degree of Correlation
(4) Low Degree of Correlation
(5) Absence of Correlation.

(1) Perfect Correlation: When two variables vary at constant ratio in the same direction, it is perfect positive correlation and when the direction of change is opposite, it is perfect negative correlation. In case of perfect positive correlation, correlation coefficient (r) is equal to +1, and in case of perfect negative correlation, correlation coefficient (r) is equal to −1.

(2) High Degree of Correlation: When correlation exists in very large magnitude, then it is called high degree of correlation. In such a case, correlation coefficient ranges between ±0.75 and ±1.

(3) Moderate Degree of Correlation: Correlation coefficient, on being within the limits of ±0.25 and ±0.75 is termed as moderate degree of correlation.

(4) Low Degree of Correlation: When correlation exists in very small magnitude, then it is called as low degree of correlation. In such a case, correlation coefficient ranges between 0 and ±0.25.

(5) Absence of Correlation: When there is no relationship between the variables, then correlation is found to be absent. In case of absence of correlation, the value of correlation coefficient is zero.

The degree of correlation on the basis of value of correlation coefficient can be summarised with the following table:

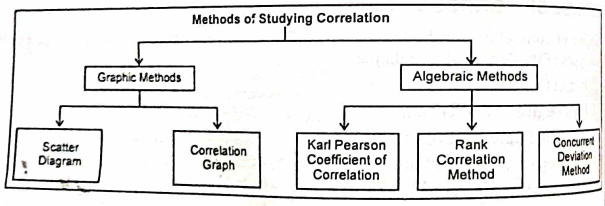
S.No.	Degree of Correlation	Positive	Negative
1.	Perfect Correlation	+1	−1
2.	High Degree of Correlation	Between +0.75 to +1	Between −0.75 to −1
3.	Moderate Degree of Correlation	Between +0.25 to +0.75	Between −0.25 to −0.75
4.	Low Degree of Correlation	Between 0 to + 0.25	Between 0 to −0.25
5.	Absence of Correlation	0	0

◘ METHODS OF STUDYING CORRELATION

Correlation can be determined by the following methods:

(1) Graphic Methods
- (*i*) Scatter Diagram
- (*ii*) Correlation Graph

(2) Algebraic Methods
- (*i*) Karl Pearson's Coefficient of Correlation
- (*ii*) Spearman's Rank Correlation Method
- (*iii*) Concurrent Deviation Method



◘ (1) GRAPHIC METHOD

• (i) Scatter Diagram

Scatter diagram is a graphic method of finding out correlation between two variables. By this method, direction of correlation can be ascertained. For constructing a scatter diagram, X-variable is represented on X-axis and the Y-variable on Y-axis. Each pair of values of X and Y series is plotted in two-dimensional space of X—Y. Thus we get a scatter diagram by plotting all the pair of values. Different points may be scattered in various ways in the scatter diagram whose analysis gives us an idea about the direction and magnitude of correlation in the following ways:

(*i*) **Perfect Positive Correlation ($r = +1$):** If all points are plotted in the shape of a straight line, passing from the lower corner of left side to the upper corner at right side, then both series X and Y have perfect positive correlation, as is clear from the diagram (A) below.

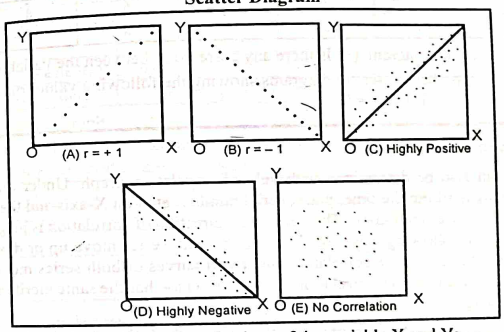
(*ii*) **Perfect Negative Correlation ($r = -1$):** When all points lie on a straight line from up to down, then X and Y have perfect negative correlation, as is clear from the diagram (B) below.

(*iii*) **High Degree of Positive Correlation:** When concentration of points moves from left to right upward and the points are close to each other, then X and Y have high degree of positive correlation, as is clear from the diagram (C) below.

(*iv*) **High Degree of Negative Correlation:** When points are concentrated from left to right downward, and the points are close to each other, then X and Y have high degree of negative correlation, as is clear from the diagram (D) below.

(*v*) **Zero Correlation ($r = 0$):** When all the points are scattered in four directions here and there and are lacking in any pattern, then there is absence of correlation, as is clear from the diagram (E) below.

Scatter Diagram

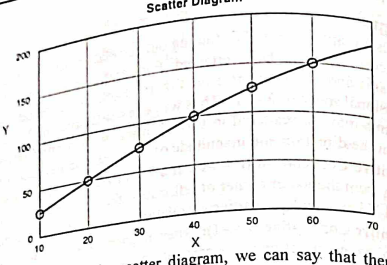


Example 1. Given the following pairs of values of the variable X and Y:

X :	10	20	30	40	50	60
Y :	25	50	75	100	125	150

(*i*) Make a Scatter Diagram.
(*ii*) Is there any correlation between the variables X and Y?

Solution: (*i*)



(*ii*) By looking at the scatter diagram, we can say that there is perfect positive correlation between X and Y variables.

▶ **Merits and Demerits of Scatter Diagram**

Determining correlation by the method is easy because no mathematical computations are to be done. The major shortcoming of this method is that degree of correlation cannot be determined.

EXERCISE 1.1

1. Given the following pairs of values of the variables X and Y:

X:	2	3	5	6	8	9
Y:	6	5	7	8	12	11

(*a*) Make a scatter diagram. (*b*) Is there any correlation between the variables X and Y?

2. Draw three hypothetical scatter diagrams showing the following values of '*r*':
(*i*) r = –1 (*ii*) r = +1 (*iii*) r = 0

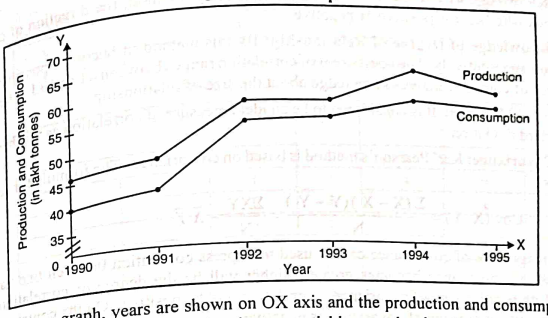
○ (ii) Correlation Graph

Correlation can also be determined with help of correlation graph. Under this method, two curves are drawn by marking the time, place, serial number, etc., on X-axis and the values of both the correlated variables' series on Y-axis. The degree and direction of correlation is judged on the basis of these curves in the following ways: (*a*) If curves of both series move up or down in the same direction, then they have positive correlation, and (b) If curves of both series move in a opposite direction, then they have negative correlation. This method too has the same merits and demerits as those of a scatter diagram.

Example 2: Construct a correlation graph on the basis of the following data and comment on the relationship between production and consumption:

Year:	1990	1991	1992	1993	1994	1995
Production (in lakh tons):	46	48	58	58	64	60
Consumption (in lakh tons):	40	42	54	55	58	57

Solution:



In above shown graph, years are shown on OX axis and the production and consumption are shown on OY axis. This graph reveals that the two variables are closely related. Both curves are moving in one direction only. The distance between them also remains almost constant, therefore, there is high degree of positive correlation between them.

EXERCISE 1.2

1. From the following data, ascertain whether the income and expenditure of the 100 workers of a factory are correlated:

Year:	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Average income (in Rs.):	100	102	105	105	101	112	118	120	125	130
Average expenditure:	90	91	93	95	92	94	100	105	108	110

Use Correlation graph. [Ans. Closely Related]

2. From the following data, ascertain with the help of correlation graph, whether the demand and price of a commodity are correlated.

Year:	1986	1987	1988	1989	1990	1991	1992	1993
Demand in units:	50	55	62	70	75	78	80	82
Price in Rs.:	40	38	35	30	27	22	20	16

[Ans. Negatively correlated]

◘ (2) ALGEBRAIC METHOD

○ (i) Karl Pearson's Coefficient of Correlation

It is quantitative method of measuring correlation. This method has been given by Karl Pearson and after his name, it is known as Pearson's coefficient of correlation. This is the best method of working out correlation coefficient. This method has the following main characteristics:

(1) Knowledge of Direction of Correlation: By this method, the direction of correlation determined whether it is positive or negative.

(2) Knowledge of Degree of Relationship: By this method, it becomes possible to measure correlation quantitatively. The coefficient of correlation ranges between –1 and +1. The value of coefficient of correlation gives knowledge about the size of relationship.

(3) Ideal Measure: It is considered to be an ideal measure of correlation as it is based on the and standard deviation.

(4) Covariance: Karl Pearson's method is based on co-variance. The formula for co-variance as follows:

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

The magnitude of co-variance can be used to express correlation between two variables. magnitude of co-variance becomes greater, higher will be the degree of correlation, other lower. With positive sign of covariance, correlation will be positive. On the contrary, correla will be negative if the sign of covariance is negative.

Calculation of Karl Pearson's Coefficient of Correlation

The calculation of Karl Pearson's coefficient of correlation can be divided into two parts:

(A) Calculation of Coefficient of Correlation in the case of Individual Series or Ungrouped Da

(B) Calculation of Coefficient of Correlation in the case of Grouped Data.

▶ (A) Calculation of Coefficient of Correlation in case of Individual Series or Ungrouped Da

The following are the main methods of calculating the coefficient of correlation in individual series:

(1) Actual Mean Method

This method is useful when arithmetic mean happens to be in whole numbers or integers. T method involves the following steps:

(1) First, we compute the arithmetic mean of X and Y series, *i.e.*, $\bar{X}$ and $\bar{Y}$ are worked out.

(2) Then from the arithmetic means of the two series, deviations of the individual values taken. The deviations of *X*-series are denoted by x and of the *Y*-series by y, *i.e.*, $x = X - \bar{X}$, $y = Y - \bar{Y}$.

(3) Deviations of the two series are squared and added up to get Σx^2 and Σy^2.

(4) The corresponding deviations of the two series (x and y) are multiplied and summed up to Σxy.

(5) Finally, correlation coefficient is found out by using the following formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \quad \text{or} \quad \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{\Sigma(Y-\bar{Y})^2}}$$

The correlation coefficient has the value always ranging between –1 and +1. The follo examples clarify the computation procedure of this method:

Example 3. From the following data, calculate Karl Pearson's coefficient of correlation:

X:	2	3	4	5	6	7	8
Y:	4	7	8	9	10	14	18

Solution:

Calculation of Coefficient of Correlation

X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2	xy
2	–3	9	4	–6	36	+18
3	–2	4	7	–3	9	+6
4	–1	1	8	–2	4	+2
5	0	0	9	–1	1	0
6	+1	1	10	0	0	0
7	+2	4	14	+4	16	+8
8	+3	9	18	+8	64	+24
$\Sigma X = 35$ $N = 7$	$\Sigma x = 0$	$\Sigma x^2 = 28$	$\Sigma Y = 70$	$\Sigma y = 0$	$\Sigma y^2 = 130$	$\Sigma xy = 58$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{35}{7} = 5, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{70}{7} = 10$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{58}{\sqrt{28 \times 130}} = \frac{58}{\sqrt{3640}}$$

$$= \frac{58}{60.33} = +0.96$$

Thus, there is a high degree of positive correlation between the variables X and Y.

Example 4. From the following data, compute the coefficient of correlation between X and Y series.

	X-Series	Y-Series
Number of items:	15	15
Arithmetic mean:	25	18
Squares of deviations from mean:	136	138

Summation of product of deviations of X and Y series from their respective arithmetic means = 122.

Solution: We are given: $N = 15$, $\bar{X} = 25$, $\bar{Y} = 18$, $\Sigma x^2 = 136$, $\Sigma y^2 = 138$, $\Sigma xy = 122$

Applying the formula,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} = \frac{122}{\sqrt{136 \times 138}}$$

$$= \frac{122}{\sqrt{18768}} = \frac{122}{136.996} = +0.89$$

IMPORTANT TYPICAL EXAMPLES

Example 5. From the following table, calculate the coefficient of correlation by Karl Pearson's method:

X:	6	2	10	4	8
Y:	9	11	–	8	7

Solution: Arithmetic means of X and Y series are 6 and 8 respectively.

Let us first find the missing value of Y and let us denote it by a.

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{9+11+a+8+7}{5} = \frac{35+a}{5}$$

$$\Rightarrow \quad 8 = \frac{35+a}{5}$$

$$\therefore \quad 35 + a = 40 \quad \Rightarrow \quad a = 5$$

Thus, the complete series is:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Now we find the coefficient of correlation.

Calculation of Coefficient of Correlation

X	$\bar{X}=6$ x	x^2	Y	$\bar{Y}=8$ y	y^2	xy
6	0	0	9	1	1	0
2	−4	16	11	3	9	−12
10	4	16	5	−3	9	−12
4	−2	4	8	0	0	0
8	2	4	7	−1	1	−2
$\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{5} = 8$$

Applying the formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-26}{\sqrt{40 \times 20}}$$

$$= \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843}$$

$$= -0.9192$$

Example 6. From the data given below, find the number of items (N):

$r = 0.5$, $\Sigma xy = 120$, Standard Deviation of Y (σ_y) = 8, $\Sigma x^2 = 90$

Where, x and y are deviations from arithmetic means.

Solution: Given: $r = 0.5$, $\Sigma xy = 120$, $\Sigma x^2 = 90$, $\sigma_y = 8$

Now, $$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} \quad \text{when } y = Y - \bar{Y} \text{ [Formula of S.D.]}$$

$$8 = \sqrt{\frac{\Sigma y^2}{N}}, \text{ squaring both sides, we get}$$

$$64 = \frac{\Sigma y^2}{N} \quad \Rightarrow \Sigma y^2 = 64N$$

Now, $$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} \quad \Rightarrow 0.5 = \frac{120}{\sqrt{90 \times 64N}}$$

Squaring both sides

$$0.25 = \frac{(120)^2}{90 \times 64N} \quad \Rightarrow 0.25 = \frac{14400}{5760N}$$

$$\Rightarrow \quad (0.25)(5760)N = 14400$$

$$\Rightarrow \quad (1440)N = 14400$$

$$\therefore \quad N = \frac{14400}{1440} = 10$$

EXERCISE 1.3

1. Calculate Karl Pearson's coefficient of correlation between the heights of fathers and sons from the following:

Height of fathers (in inches):	65	66	67	68	69	70	71
Height of sons (in inches):	67	68	66	69	72	72	69

[Ans. $r = 0.668$]

2. Calculate Pearson's coefficient of correlation between X and Y from the following data:

X:	14	19	24	21	26	22	15	20	19
Y:	31	36	48	37	50	45	33	41	39

[Ans. $r = 0.947$]

3. Calculate the coefficient of correlation using Karl Pearson's formula based on actual mean value of the series given below:

X:	10	12	15	23	20
Y:	14	17	23	25	21

[Ans. $r = 0.864$]

4. From the following data, compute Karl Pearson coefficient of correlation:

	X-series	Y-series
Number of items:	7	7
Arithmetic mean:	4	8
Sum of squares of deviations from arithmetic mean:	28	76
Summation of products of deviations of X and Y series from their respective means is		

[Ans. r =

5. If $r = 0.25$, $\Sigma xy = 45$, $\sigma_y = 3$, $\Sigma x^2 = 50$, where x and y denote deviations from their respective means, find the number of observations. [Ans. N

6. Two variates X and Y when expressed as deviations from their respective means are given as follows:

x:	–4	–3	–1	–2	0	1	2	3	4
y:	3	–3	?	0	4	1	2	–2	–1

Find the coefficient of correlation between them. [Ans. r =

[**Hint:** See Example 51]

7. Calculate Karl Pearson's coefficient of correlation, taking deviations from actual means 52 and 44 of the following data:

X:	44	46	46	48	52	54	?	56	60	8
Y:	36	40	42	40	?	44	46	48	50	1

[Ans. r = +0.9

8. Determine Pearson's coefficient of correlation from the following data:
$\Sigma X = 250, \Sigma Y = 300, N = 10, \Sigma(X-25)^2 = 480, \Sigma(Y-30)^2 = 600$ and $\Sigma(X-25)(Y-30) = 150$ [Ans. r =

(2) Assumed Mean Method

This method is useful when arithmetic mean is not in whole numbers but in fractions. In this method, deviations from assumed means of both the series (X and Y) are calculated. Correlation coefficient by this method can be determined in the following manner:

(1) Any values of X and Y are taken as their assumed mean, Ax and Ay.

(2) Deviations of the individual values of both the series (X and Y) are worked out from their assumed means. Deviations of X series ($X - Ax$) are denoted by dx and of Y series ($Y - Ay$) by dy.

(3) Deviations are summed up to get Σdx and Σdy.

(4) Then, squares of the deviations dx^2 and dy^2 are worked out and summed up to get Σdx^2 and Σdy^2 respectively.

(5) Each dx is multiplied by the corresponding dy and the products ($dxdy$) are added up to get $\Sigma dxdy$.

(6) Finally, correlation coefficient is obtained by using any one of following formula:

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx \cdot \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}} \quad ...(i)$$

or

$$r = \frac{N \cdot \Sigma dxdy - \Sigma dx \cdot \Sigma dy}{\sqrt{N \cdot \Sigma dx^2 - (\Sigma dx)^2}\cdot\sqrt{N \cdot \Sigma dy^2 - (\Sigma dy)^2}} \quad ...(ii)$$

or

$$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N \cdot \sigma_x \cdot \sigma_y} \quad ...(iii)$$

Note: Unless otherwise specifically asked, formula (*ii*) should be used as it makes the computation work very easy.

The following examples clarify the computation process of this method:

Example 1. Find the coefficient of correlation from the following data:

X:	10	12	18	16	15	19	18	17
Y:	30	35	45	44	42	48	47	46

Solution:

Calculation of Coefficient of Correlation

X	A=16 dx	dx^2	Y	A=42 dy	dy^2	$dxdy$
10	–6	36	30	–12	144	72
12	–4	16	35	–7	49	28
18	+2	4	45	+3	9	6
16 = A	0	0	44	+2	4	0
15	–1	1	42 = A	0	0	0
19	+3	9	48	+6	36	18
18	+2	4	47	+5	25	10
17	+1	1	46	+4	16	4
$\Sigma X = 125$ $N = 8$	$\Sigma dx = -3$	$\Sigma dx^2 = 71$	$\Sigma Y = 337$ $N = 8$	$\Sigma dy = 1$	$\Sigma dy^2 = 283$	$\Sigma dxdy = 138$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{125}{8} = 15.62 \quad , \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{337}{8} = 42.12$$

Since the actual means are not whole numbers, we take 16 as assumed mean for X and 42 as assumed mean for Y.

Applying the formula,

$$r = \frac{N \cdot \Sigma dxdy - \Sigma dx \cdot \Sigma dy}{\sqrt{N \cdot \Sigma dx^2 - (\Sigma dx)^2}\cdot\sqrt{N \cdot \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8 \times 138 - (-3)(1)}{\sqrt{8 \times 71 - (-3)^2}\sqrt{8 \times 283 - (1)^2}}$$

$$= \frac{1104+3}{\sqrt{568-9}\ \sqrt{2264-1}} = \frac{1107}{\sqrt{559}\ \sqrt{2263}}$$

$$= \frac{1107}{\sqrt{1265017}} = \frac{1107}{1124.72} = 0.98$$

Aliter: $\bar{X} = 15.62,\ \bar{Y} = 42.12,\ Ax = 16,\ Ay = 42,\ \Sigma dxdy = 138$

$$\sigma_x = \sqrt{\frac{\Sigma dx^2}{N} - \left(\frac{\Sigma dx}{N}\right)^2} = \sqrt{\frac{71}{8} - \left(\frac{-3}{8}\right)^2} = \sqrt{\frac{71}{8} - \frac{9}{64}} = 2.95$$

$$\sigma_y = \sqrt{\frac{\Sigma dy^2}{N} - \left(\frac{\Sigma dy}{N}\right)^2} = \sqrt{\frac{283}{8} - \left(\frac{1}{8}\right)^2} = \sqrt{\frac{283}{8} - \frac{1}{64}} = 5.94$$

Applying the formula

$$r = \frac{\Sigma dxdy - N(\bar{X} - Ax)(\bar{Y} - Ay)}{N.\sigma_x.\sigma_y}$$

$$= \frac{138 - 8(15.62 - 16)\,(42.12 - 42)}{8 \times 2.95 \times 5.94} = \frac{138 - 8(-0.38)(0.12)}{140.184}$$

$$= \frac{138 - 0.3648}{140.184} = \frac{137.6352}{140.184} = 0.98$$

Example 8. Calculate Karl Pearson's coefficient of correlation from the following data:

X:	24	27	28	28	29	30	32	33	35	35	40
Y:	18	20	22	25	22	28	28	30	27	30	32

(You may use 32 as working mean for X and 25 that for Y.)

Solution:

Calculation of Coefficient of Correlation

X	A=32 dx	dx^2	Y	A=25 dy	dy^2	$dxdy$
24	−8	64	18	−7	49	56
27	−5	25	20	−5	25	25
28	−4	16	22	−3	9	12
28	−4	16	25 = A	0	0	0
29	−3	9	22	−3	9	9
30	−2	4	28	+3	9	−6
32 = A	0	0	28	+3	9	0
33	1	1	30	+5	25	5
35	3	9	27	+2	4	6
35	3	9	30	+5	25	15
40	8	64	22	−3	9	−24
$N = 11$	$\Sigma dx = -11$	$\Sigma dx^2 = 217$		$\Sigma dy = -3$	$\Sigma dy^2 = 173$	$\Sigma dxdy$

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx\,.\,\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\ \sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}}$$

$$= \frac{98 - \frac{(-11)(-3)}{11}}{\sqrt{217 - \frac{(-11)^2}{11}} \times \sqrt{173 - \frac{(-3)^2}{11}}} = \frac{98 - 3}{\sqrt{217 - 11}\ \sqrt{173 - 0.82}}$$

$$= \frac{95}{\sqrt{206 \times 172.18}} = \frac{95}{188.33} = 0.504$$

Aliter: r can be calculated by using the formula:

$$r = \frac{\Sigma dxdy \times N - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{98 \times 11 - (-11)(-3)}{\sqrt{11 \times 217 - (-11)^2} \times \sqrt{11 \times 173 - (-3)^2}}$$

$$= \frac{1078 - 33}{\sqrt{2387 - 121}\ \sqrt{1903 - 9}} = \frac{1045}{\sqrt{2266}\ \sqrt{1894}}$$

$$= \frac{1045}{\sqrt{4291804}} = \frac{1045}{2071.66} = 0.504$$

Example 9. Deviations of the items of two series X and Y from assumed mean are as under:

Deviations of X:	+5	−4	−2	+20	−10	0	+3	0	−15	−5
Deviations of Y:	+5	−12	−7	+25	−10	−3	0	+2	−9	−15

Calculate Karl Pearson's coefficient of correlation.

Solution:

dx	dx^2	dy	dy^2	$dxdy$
+5	25	+5	25	25
−4	16	−12	144	48
−2	4	−7	49	14
+20	400	+25	625	500
−10	100	−10	100	100
0	0	−3	9	0
+3	9	0	0	0
0	0	+2	4	0
−15	225	−9	81	135
−5	25	−15	225	75
$\Sigma dx = -8$	$\Sigma dx^2 = 804$	$\Sigma dy = -24$	$\Sigma dy^2 = 1262$	$\Sigma dxdy = 897$

$$r = \frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{10 \times 897 - (-8)(-24)}{\sqrt{10 \times 804 - (-8)^2}\ \sqrt{10 \times 1262 - (-24)^2}}$$

$$= \frac{8970 - 192}{\sqrt{8040 - 64}\ \sqrt{12620 - 576}} = \frac{8778}{\sqrt{7976}\ \sqrt{12044}}$$

$$= \frac{8778}{\sqrt{96062944}} = \frac{8778}{9801.17} = 0.895$$

Calculation of Coefficient of Correlation by taking a Common Factor

Common factor may be used to simplify the calculation of coefficient of correlation. It is important to note here that there will be no effects on the formula of coefficient of correlation if common factor is used. The main reason is that the coefficient of correlation is independent of change of origin and scale. If the origin is shifted or scale is changed, it will not affect the value of coefficient of correlation.

Example 10. Calculate coefficient of correlation from the following data:

X:	100	200	300	400	500	60
Y:	110	120	135	140	160	16

Solution: To simplify the calculation, let

$$dx = \frac{X - 400}{100},\ dy = \frac{Y - 140}{5}$$

Calculation of Coefficient of Correlation

X	dx	dx^2	Y	dy	dy^2	dx
100	–3	9	110	–6	36	II
200	–2	4	120	–4	16	II
300	–1	1	135	–1	1	I
400	0	0	140	0	0	I
500	+1	1	160	4	16	I
600	+2	4	165	5	25	II
$N = 6$	$\Sigma dx = -3$	$\Sigma dx^2 = 19$		$\Sigma dy = -2$	$\Sigma dy^2 = 94$	Σdx

$$r = \frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6 \times 41 - (-3)(-2)}{\sqrt{6 \times 19 - (-3)^2} \times \sqrt{6 \times 94 - (-2)^2}}$$

$$= \frac{246 - 6}{\sqrt{105}\ \sqrt{560}} = \frac{240}{\sqrt{58800}} = \frac{240}{242.487} = 0.9897$$

IMPORTANT TYPICAL EXAMPLES

Example 11. From the following data, calculate the Karl Pearson's coefficient of correlation between age of students and their playing habits:

Age:	15	16	17	18	19	20
No. of students:	250	200	150	120	100	80
Regular players:	200	150	90	48	30	12

Solution: Since it is asked to find the correlation between age and playing habits, it is required to find the percentage of regular players which is obtained as follows:

No. of students	Regular players	% of Regular players
250	200	$\frac{200}{250} \times 100 = 80$
200	150	$\frac{150}{200} \times 100 = 75$
150	90	$\frac{90}{150} \times 100 = 60$
120	48	$\frac{48}{120} \times 100 = 40$
100	30	$\frac{30}{100} \times 100 = 30$
80	12	$\frac{12}{80} \times 100 = 15$

Now we calculate the correlation coefficient between age and percentage of regular players. Denoting the age by X and percentage of regular players by Y.

X	dx	dx^2	Y	dy	dy^2	$dxdy$
15	–2	4	80	+20	400	–40
16	–1	1	75	+15	225	–15
17 = A	0	0	60 = A	0	0	0
18	+1	1	40	–20	400	–20
19	+2	4	30	–30	900	–60
20	+3	9	15	–45	2025	–135
N = 6	$\Sigma dx = 3$	$\Sigma dx^2 = 19$		$\Sigma dy = -60$	$\Sigma dy^2 = 3950$	$\Sigma dxdy = -270$

$$r = \frac{N \times \Sigma dxdy - \Sigma dx\ \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6 \times (-270) - (3)(-60)}{\sqrt{6 \times 19 - (3)^2} \times \sqrt{6 \times 3950 - (-60)^2}}$$

$$= \frac{-1620+180}{\sqrt{114-9}\sqrt{23700-3600}} = \frac{-1440}{\sqrt{2110500}} = \frac{-1440}{1452.75} = -0.99$$

There is a high degree of negative correlation between age and playing habits. It shows that as age increases, the tendency to play decreases.

Example 12. From the following data, calculate Karl Pearson's coefficient of correlation between age and blindness:

Age	No. of persons (in thousands)	Blinds
0—10	100	55
10—20	60	40
20—30	40	40
30—40	36	40
40—50	24	36
50—60	11	22
60—70	6	18
70—80	3	15

Solution: First, we shall find the number of blinds per lakh of population in each group as:

No. of persons ('000)	Blinds	No. of blinds (per lakh)
100	55	$\frac{55}{100000}\times 100000 = 55$
60	40	$\frac{40}{60000}\times 100000 = 67$
40	40	$\frac{40}{40000}\times 100000 = 100$
36	40	$\frac{40}{36000}\times 100000 = 111$
24	36	$\frac{36}{24000}\times 100000 = 150$
11	22	$\frac{22}{11000}\times 100000 = 200$
6	18	$\frac{18}{6000}\times 100000 = 300$
3	15	$\frac{15}{3000}\times 100000 = 500$

Denoting the Mid Value of Age by X and No. of Blinds per lakh by Y, we find coefficient of correlation.

Age	MV (X)	A = 35, $dx = \frac{X-35}{10}$	dx^2	Y	A=185, $dy = Y-185$	dy^2	$dxdy$
0—10	5	−3	9	55	−130	16900	390
10—20	15	−2	4	67	−118	13924	236
20—30	25	−1	1	100	−85	7225	85
30—40	35	0	0	111	−74	5476	0
40—50	45	+1	1	150	−35	1225	−35
50—60	55	+2	4	200	+15	225	30
60—70	65	+3	9	300	+115	13225	345
70—80	75	+4	16	500	+315	99225	1260
$N = 8$		$\Sigma dx = 4$	$\Sigma dx^2 = 44$		$\Sigma dy = 3$	$\Sigma dy^2 = 157425$	$\Sigma dxdy = 2311$

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx \cdot \Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{(\Sigma dx)^2}{N}}\sqrt{\Sigma dy^2 - \frac{(\Sigma dy)^2}{N}}} = \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}}$$

$$= \frac{2311 - \frac{(4)(3)}{8}}{\sqrt{44 - \frac{(4)^2}{8}}\sqrt{157425 - \frac{(3)^2}{8}}} = \frac{2311 - 1.5}{\sqrt{42}\sqrt{157423.87}}$$

$$= \frac{2309.5}{\sqrt{42}\sqrt{157423.87}} = \frac{2309.5}{6.48 \times 396.76} = \frac{2309.5}{2571.004} = +0.898$$

Example 13. From the following data, calculate the coefficient of correlation between X-series and Y-series.

	X-series	Y-series
Mean	74.5	125.5
Assumed mean	69	112
Standard deviation (σ)	13.07	15.85

Sum of products of corresponding deviations of X and Y series from their assumed mean ($\Sigma dxdy$) = 2176 and no. of pairs of observations = 8.

Solution: Given:

$N = 8$, $\bar{X} = 74.5$, $A_x = 69$, $\sigma_x = 13.07$,

$\bar{Y} = 125.5$, $A_y = 112$, $\sigma_y = 15.85$, $\Sigma dxdy = 2176$

Applying the formula:

$$r = \frac{\Sigma dxdy - N(\bar{X} - A_x)(\bar{Y} - A_y)}{N \cdot \sigma_x \cdot \sigma_y}$$

Substituting the values in the formula:

$$= \frac{2176 - 8(74.5 - 69)(125.5 - 112)}{8 \times 13.07 \times 15.85}$$

$$= \frac{2176 - 8(5.5)(13.5)}{8 \times 13.07 \times 15.85} = \frac{2176 - 594}{1657.276} = \frac{1582}{1657.276} = +0.954$$

Example 14. From the following data, calculate the coefficient of correlation between 'age' and 'playing habits':

Age	No. of students	No. of regular players
15—16	200	150
16—17	270	162
17—18	340	170
18—19	360	180
19—20	400	180
20—21	300	120

Solution: First we shall find the percentage of regular players as follows:

No. of students	No. of regular players	% of regular players
200	150	$\frac{150}{200} \times 100 = 75$
270	162	$\frac{162}{270} \times 100 = 60$
340	170	$\frac{170}{340} \times 100 = 50$
360	180	$\frac{180}{360} \times 100 = 50$
400	180	$\frac{180}{400} \times 100 = 45$
300	120	$\frac{120}{300} \times 100 = 40$

Denoting Mid-Value of Age by X and Percentage of Regular Players by Y.

Calculation of Coefficient of Correlation

Age	M.V. (X)	A=17.5 dx	dx^2	% of Regular players (Y)	A= 50 dy	dy^2	$dxdy$
15–16	15.5	−2	4	75	+25	625	−50
16–17	16.5	−1	1	60	+10	100	−10
17–18	17.5 = A	0	0	50 = A	0	0	0
18–19	18.5	+1	1	50	0	0	0
19–20	19.5	+2	4	45	−5	25	−10
20–21	20.5	+3	9	40	−10	100	−30
$N = 6$		$\Sigma dx = 3$	$\Sigma dx^2 = 19$		$\Sigma dy = 20$	$\Sigma dy^2 = 850$	$\Sigma dxdy = -$

Now,

$$r = \frac{N \times \Sigma dxdy - \Sigma dx\, \Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6 \times (-100) - (3)(20)}{\sqrt{6 \times 19 - (3)^2}\ \sqrt{6 \times 850 - (20)^2}}$$

$$= \frac{-660}{\sqrt{105}\ \sqrt{4700}} = \frac{-660}{702.4956} = -0.9395$$

It shows that there is high degree of negative correlation between age and playing habits.

Example 15. From the data given below, calculate the coefficient of correlation by Karl Pearson's method between density of population and death rate:

Cities	Area in sq. miles	Population (in '000)	No. of deaths
A	150	30	300
B	180	90	1440
C	100	40	560
D	60	42	840
E	120	72	1224
F	80	24	312

Solution: First we calculate density of population and death rate by using the formula and denote them by X and Y.

$$\text{Density of Population} = \frac{\text{Population}}{\text{Area}}$$

$$\text{Death Rate} = \frac{\text{No. of Deaths}}{\text{Population}} \times 1000$$

Formula.

Cities	Area (in sq. mile)	Population ('000)	No. of deaths	Density (X)	Death rate (%) (Y)
A	150	30,000	300	$\frac{30,000}{150} = 200$	$\frac{300}{30,000} \times 1,000 = 10$
B	180	90,000	1440	$\frac{90,000}{180} = 500$	$\frac{1,440}{90,000} \times 1,000 = 16$
C	100	40,000	560	$\frac{40,000}{100} = 400$	$\frac{560}{40,000} \times 1,000 = 14$
D	60	42,000	840	$\frac{42,000}{60} = 700$	$\frac{840}{42,000} \times 1,000 = 20$
E	120	72,000	1224	$\frac{72,000}{120} = 600$	$\frac{1224}{72,000} \times 1,000 = 17$
F	80	24,000	312	$\frac{24,000}{80} = 300$	$\frac{312}{24,000} \times 1,000 = 13$

Calculation of Coefficient of Correlation

Cities	Density (X)	$\bar{X}=450$, $x=\frac{X-450}{50}$	x^2	Death Rate (Y)	$\bar{Y}=15$, $y=Y-15$	y^2	xy
A	200	−5	25	10	−5	25	25
B	500	+1	1	16	+1	1	1
C	400	−1	1	14	−1	1	1
D	700	+5	25	20	+5	25	25
E	600	+3	9	17	+2	4	6
F	300	−3	9	13	−2	4	6
$N=6$	$\Sigma X=2700$	$\Sigma x=0$	$\Sigma x^2=70$	$\Sigma Y=90$	$\Sigma y=0$	$\Sigma y^2=60$	$\Sigma xy=64$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{2700}{6}=450, \qquad \bar{Y}=\frac{\Sigma Y}{N}=\frac{90}{6}=15$$

Since the actual means of X and Y are whole numbers, we should take deviations from actual means of X and Y to simplify the calculations:

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

$$=\frac{64}{\sqrt{70}\times\sqrt{60}}=\frac{64}{\sqrt{70\times 60}}$$

$$=\frac{64}{64.81}=+0.9875$$

There is a high degree of positive correlation between density of population and death rate.

EXERCISE 1.4

1. Calculate the Correlation Coefficient from the following data of marks obtained in Commerce (X) and Economics (Y):

X:	50	60	58	47	49	33	65	43	46	61
Y:	48	65	50	48	55	58	63	48	50	71

[Ans. r=0.6...]

2. Seven students obtained the following percentage of marks in the college test (X) and in the final examination (Y). Find out the coefficient of correlation between these variables:

X:	50	62	72	25	20	60	68
Y:	48	65	74	33	25	55	68

[Ans. r=0...]

3. Calculate Karl Pearson's coefficient of correlation between the values of X and Y for the following data:

X:	78	89	96	69	59	79	68	61
Y:	125	137	156	112	107	136	123	108

Assume 69 and 112 as the mean values for X and Y respectively. [Ans. r=+0.954]

4. From the following data, calculate the coefficient of correlation between X-series and Y-series:

	X-series	Y-series
Mean:	381.2	24.5
Assumed mean:	380	25
Standard deviation (σ):	16.79	2.97

Summation of products of corresponding deviations of X and Y series from their assumed means ($\Sigma dxdy$) = 390 and no. of pairs of observations = 10. [Ans. r=0.794]

5. The following table gives the distribution of items of production and also the relative defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality.

Size group:	15—16	16—17	17—18	18—19	19—20	20—21
No. of items:	200	270	340	360	400	300
No. of defective items:	150	162	170	180	180	114

[**Hint:** See Example 52] [Ans. r=−0.95]

6. Find out coefficient of correlation from the following data:

X:	300	350	400	450	500	550	600	650	700
Y:	800	900	1000	1100	1200	1300	1400	1500	1600

[**Hint:** Let $dx=\frac{X-500}{50}$, $dy=\frac{Y-1200}{100}$] [Ans. r=+1]

7. Calculate the coefficient of correlation between age group and mortality rate from the following data :

Age group :	0—20	20—40	40—60	60—80	80—100
Rate of mortality :	350	280	540	760	900

[Ans. r = 0.947]

8. Calculate Karl Pearson's coefficient of correlation between age and playing habits from the data given below:

Age:	16	17	18	19	20	21	22
No. of students:	350	320	280	240	180	120	50
Regular players:	315	256	182	132	63	18	4

[Ans. r = −0.994]

9. Following figures give the rainfall in inches and production in '00 tons for Rabi and Kharif crops for number of years. Find the coefficient of correlation between rainfall and production:

Rainfall:	20	22	24	26	28	30	32
Rabi production:	15	18	20	32	40	39	45
Kharif production:	15	17	20	18	20	21	15

[Ans. $r = +0.7$]

10. With the following data in 4 cities, calculate the coefficient of correlation by Pearson's method between the density of population and the death rate:

Cities	Area in sq.km.	Population ('000)	No. of deaths
A	200	40	480
B	150	75	1200
C	120	72	1080
D	80	20	280

[Ans. $r = +0.9$]

11. Calculate 'r' from the following data:
$\Sigma X = 225, \Sigma Y = 189, N = 10, \Sigma(X-22)^2 = 85, \Sigma(Y-19)^2 = 25$ and $\Sigma(X-22)(Y-19) = 45$
[Hint: See Example 53 Aliter] [Ans. $r = 0.9$]

(3) Method Based on the Use of Actual Data

This method is also known as **Product moment method**. When number of observations is few, correlation coefficient can also be calculated without taking deviations either from actual mean or from assumed mean *i.e.* from actual X and Y values. In this method, the correlation coefficient can be determined in the following way:

(1) First of all, values of the variables X and Y series are summed up to get ΣX and ΣY.

(2) The values of the variables of X and Y series are squared up and added to get ΣX^2 and ΣY^2.

(3) The values of X variable and Y variable are multiplied and the product is added up to get ΣXY.

(4) Finally, the following formula is used to get the correlation coefficient:

$$r = \frac{\Sigma XY - \frac{\Sigma X \cdot \Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}} \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

or

$$r = \frac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{N \cdot \Sigma X^2 - (\Sigma X)^2} \sqrt{N \cdot \Sigma Y^2 - (\Sigma Y)^2}}$$

Example 16. From the following data, find Karl Pearson coefficient of correlation:

X:	2	3	1	5	6	4
Y:	4	5	3	4	6	2

Solution:

Calculation of Coefficient of Correlation

X	X^2	Y	Y^2	XY
2	4	4	16	8
3	9	5	25	15
1	1	3	9	3
5	25	4	16	20
6	36	6	36	36
4	16	2	4	8
$N = 6, \Sigma X = 21$	$\Sigma X^2 = 91$	$\Sigma Y = 24$	$\Sigma Y^2 = 106$	$\Sigma XY = 90$

Applying the formula:

$$r = \frac{N \cdot \Sigma XY - \Sigma X \cdot \Sigma Y}{\sqrt{N \cdot \Sigma X^2 - (\Sigma X)^2} \sqrt{N \cdot \Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{6 \times 90 - (21)(24)}{\sqrt{6 \times 91 - (21)^2} \sqrt{6 \times 106 - (24)^2}} = \frac{540 - 504}{\sqrt{546 - 441} \sqrt{636 - 576}}$$

$$= \frac{36}{\sqrt{105}\ \sqrt{60}} = \frac{0.36}{\sqrt{6300}} = \frac{36}{79.37} = +0.453$$

Example 17. Calculate product moment correlation coefficient from the following data:

X:	–5	–10	–15	–20	–25	–30
Y:	50	40	30	20	10	5

Solution: In this question the mean of X and Y series may come in fractions or negative signs. It will pose a problem in computing deviations, so here method based on the use of actual values will be used.

Calculation of Coefficient of Correlation

X	X^2	Y	Y^2	XY
–5	25	50	2500	–250
–10	100	40	1600	–400
–15	225	30	900	–450
–20	400	20	400	–400
–25	625	10	100	–250
–30	900	5	25	–150
$\Sigma X = -105$ $N = 6$	$\Sigma X^2 = 2275$	$\Sigma Y = 155$	$\Sigma Y^2 = 5525$	$\Sigma XY = -1900$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{6\times(-1900) - (-105)(155)}{\sqrt{6\times 2275 - (-105)^2}\ \sqrt{6\times 5525 - (155)^2}}$$

$$= \frac{-11400 + 16275}{\sqrt{13650 - 11025}\ \sqrt{33150 - 24025}}$$

$$= \frac{4875}{\sqrt{2625}\sqrt{9125}} = \frac{4875}{\sqrt{23953125}} = \frac{4875}{4894.19} = 0.996$$

Example 18. Find the Coefficient of Correlation for the following data:
$N = 10,\ \bar{X} = 5.5,\ \bar{Y} = 4,\ \Sigma X^2 = 385,\ \Sigma Y^2 = 192,\ \Sigma(X+Y)^2 = 947$

Solution: $\bar{X} = \frac{\Sigma X}{N} \Rightarrow 5.5 = \frac{\Sigma X}{10} \Rightarrow \Sigma X = 55$

$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow 4 = \frac{\Sigma Y}{10} \Rightarrow \Sigma Y = 40$

$$\Sigma(X+Y)^2 = \Sigma X^2 + \Sigma Y^2 + 2\Sigma XY = 947$$

$$\Rightarrow 385 + 192 + 2\Sigma XY = 947 \Rightarrow 2\Sigma XY = 370$$

$$\Rightarrow \Sigma XY = 185$$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

Putting the given values, we get

$$= \frac{10\times 185 - (55)(40)}{\sqrt{10\times 385 - (55)^2}\ \sqrt{10\times 192 - (40)^2}}$$

$$= \frac{1850 - 2200}{\sqrt{3850 - 3025}\ \sqrt{1920 - 1600}} = \frac{-350}{\sqrt{825\times 320}}$$

$$= \frac{-350}{513.80} = -0.681$$

IMPORTANT TYPICAL EXAMPLES

Example 19. Calculate the coefficient of correlation from the following data and interpret result:
$\Sigma XY = 8425, \bar{X} = 28.5, \bar{Y} = 28.0, \sigma_x = 10.5, \sigma_y = 5.6$ and $N = 10$

Solution: On the basis of informations given, we use direct method for the calculation correlation coefficient:

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

For this formula, the value of ΣXY and N are known, the values of $\Sigma X, \Sigma Y, \Sigma X^2$ and ΣY^2 are to be calculated.

$\bar{X} = \frac{\Sigma X}{N} \Rightarrow \Sigma X = N\bar{X} = 10 \times 28.5 = 285$ ...(i)

$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow \Sigma Y = N\bar{Y} = 10 \times 28.0 = 280$ ...(ii)

$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$ (Formula of S.D.)

$\Rightarrow \sigma_x^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2$

$\therefore \Sigma X^2 = N[\sigma^2{}_x + (\bar{X})^2] = 10[(10.5)^2 + (28.5)^2] = 9225$ ...(iii)

Similarly, $\Sigma Y^2 = N[\sigma_y^2 + (\bar{Y})^2] = 10[(5.6)^2 + (28.0)^2] = 8153.6$ ...(iv)

$\Sigma XY = 8425$ (given), $N = 10$

Now,
$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10\times 8425 - (285)(280)}{\sqrt{9225\times 10 - (285)^2}\sqrt{8153.6\times 10 - (280)^2}}$$

$$= \frac{4450}{\sqrt{11025}\sqrt{3136}} = \frac{4450}{5880} = 0.756$$

Interpretation: There is a positive correlation between X and Y.

Aliter: r can be calculated as follows:

$$r = \frac{Cov(X,Y)}{\sigma_x.\sigma_y}$$

$$Cov(X, Y) = \frac{1}{N}\Sigma(X - \bar{X})(Y - \bar{Y}) = \frac{1}{N}\Sigma XY - \bar{X}\bar{Y}$$

Substituting the values, we have

$$Cov(X, Y) = \frac{1}{10}(8425) - (28.5)(28.0)$$

$$= 842.5 - 798 = 44.5$$

Now,
$$r = \frac{Cov.(X,Y)}{\sigma_x.\sigma_y} = \frac{44.5}{(10.5)(5.6)} = \frac{44.5}{58.8} = 0.756$$

From the value of $r = 0.756$, it appears that there is positive correlation between X and Y.

Example 20. The following are the nine pairs of values of variable X and Y:
$N = 9,\ \Sigma X = 45,\ \Sigma Y = 135,\ \Sigma X^2 = 285,\ \Sigma Y^2 = 2085,\ \Sigma XY = 731$
While checking it was found out that two pairs were copied as:

X	Y
8	10
6	8

instead of

12	6
10	7

Obtain the correlation coefficient for the corrected data.

Solution: $N = 9,\ \Sigma X = 45,\ \Sigma Y = 135,\ \Sigma X^2 = 285,\ \Sigma Y^2 = 2085,\ \Sigma XY = 731$

Replacing the wrong values by correct values, new values are

$\Sigma X = 45 - 8 - 6 + 12 + 10 = 53$

$\Sigma Y = 135 - 10 - 8 + 6 + 7 = 130$

$\Sigma X^2 = 285 - 64 - 36 + 144 + 100 = 429$

$\Sigma Y^2 = 2085 - 100 - 64 + 36 + 49 = 2006$

$\Sigma XY = 731 - 80 - 48 + 72 + 70 = 745$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{9\times 745 - (53)(130)}{\sqrt{9\times 429 - (53)^2}\ .\sqrt{9\times 2006 - (130)^2}}$$

$$= -0.153$$

Example 21. While calculating the coefficient of correlation between the variables X and Y, a computer obtained the following constants:
$N = 20, r = 0.3, \bar{X} = 15, \bar{Y} = 20, \sigma_x = 4$ and $\sigma_y = 5$
In the course of checking, however, it was detected that an item 27 has been wrongly taken as 17 in case of X series and 35 instead of 30 in case of Y series. Obtain the correct value of r.

Solution: Given $N = 20, \bar{X} = 15, \bar{Y} = 20, \sigma_x = 4, \sigma_y = 5, r = 0.3$

We have $\bar{X} = \frac{\Sigma X}{N} \Rightarrow \Sigma X = N\bar{X} = 20\times 15 = 300$

But this is not the correct value of ΣX due to mistakes

Corrected $\Sigma X = 300 - 17 + 27 = 310$

$$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow \Sigma Y = N\bar{Y} = 20\times 20 = 400$$

But this is not the correct value of ΣY due to mistakes

Corrected $\Sigma Y = 400 - 35 + 30 = 395$

We know $\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - (\bar{X})^2}$ (Formula of S.D.)

$$\sigma_x^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2 \qquad \therefore\ \Sigma X^2 = N[\sigma_x^2 + (\bar{X})^2] = 20[16 + 225] = 4820$$

But this is not the correct value of ΣX^2 due to mistakes

Corrected $\Sigma X^2 = 4820 - 17^2 + 27^2 = 4820 - 289 + 729 = 5260$ ...(iii)

$$\sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - (\bar{Y})^2} \qquad \text{(Formula of S.D.)}$$

$$\Rightarrow \quad \sigma_y^2 = \frac{\Sigma Y^2}{N} - (\bar{Y})^2 \qquad \therefore\ \Sigma Y^2 = N[\sigma_y^2 + (\bar{Y})^2] = 20[25 + 400] = 8500$$

But this is not the correct value of ΣY^2 due to mistakes

Corrected $\Sigma Y^2 = 8500 - 35^2 + 30^2 = 8500 - 1225 + 900 = 8175$ ...(iv)

Calculation of Corrected ΣXY

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$0.3 = \frac{20\times \Sigma XY - (300)(400)}{\sqrt{20\times 4820 - (300)^2}\ \sqrt{20\times 8500 - (400)^2}}$$

$$0.3 = \frac{20\Sigma XY - 1{,}20{,}000}{80\times 100}$$

$$0.3 \times 8000 = 20\Sigma XY - 1{,}20{,}000$$

$$20\Sigma XY = 1{,}22{,}400$$

$$\Rightarrow \quad \Sigma XY = 6120$$

$\therefore$ Incorrected $\Sigma XY = 6120$

But this is not the correct value of ΣXY due to mistakes

Corrected $\Sigma XY = 6120 + 810 - 595 = 6335$ ...(v)

Now, the correct value of r would be calculated as:

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{20\times 6335 - (310)(395)}{\sqrt{20\times 5260 - (310)^2}\sqrt{20\times 8175 - (395)^2}}$$

$$= \frac{126700 - 122450}{\sqrt{105200 - 96100}\sqrt{163500 - 156025}}$$

$$= \frac{4250}{\sqrt{9100}\sqrt{7475}} = \frac{4250}{8247.57} = 0.5153$$

EXERCISE 1.5

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

X:	5	4	3	2	1
Y:	5	2	10	8	4

What will be the correlation coefficient between $2X+3$ and $5Y-4$.
[Hint: See Example 50] [Ans. $r=-0.1980$, No ...]

2. Calculate Karl Pearson's coefficient of correlation between the values of X and Y given below:

X:	−15	+18	−12	−10	+15	−20	−25	+15	+16	−16
Y:	+8	−10	+5	+12	−6	+4	+11	−9	−7	+11

[Ans. $r=-0.9$...]

3. Calculate 'r' from the following data:
$\Sigma X = 225, \Sigma Y = 189, N = 10, \Sigma(X-22)^2 = 85$
$\Sigma(Y-19)^2 = 25$ and $\Sigma(X-22)(Y-19) = 43$
[Hint: See Example 53] [Ans. $r=0.9$...]

4. Following result were obtained from an analysis of 12 pairs of observations:
$n = 12, \Sigma X = 30, \Sigma Y = 5, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 334$
Later on it was discovered that one pair of values ($X = 11, Y = 4$) were copied wrongly, ... correct values of the pair was ($X = 10, Y = 14$). Find the correct value of correlation coefficient. [Ans. $r=0.7$...]

5. Calculate the coefficient of correlation from the following data and interpret the result:
$N = 10, \bar{X} = 15, \bar{Y} = 12, \Sigma XY = 1500, \sigma_x = 4, \sigma_y = 9.0$ [Ans. $r=-0.$...]

6. Given the following:
$r = -1, \bar{X} = 4.5, \bar{Y} = 5.5, \sigma_x^2 = 5.25, \sigma_y^2 = 5.25, N = 8$
One pair of observation ($X = 9, Y = 10$) omitted to be included and hence to be incl... calculate the correct coefficient of correlation. [Ans. $r=-1$...]

7. In two sets of variables X and Y with 50 items each, the following data were observed:
$\bar{X} = 10, \sigma_x = 3, \bar{Y} = 6, \sigma_y = 2, r = 0.3$
However, on subsequent verification it was found that one value of X(=10) and one val... Y(=6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, ... the original value of correlation coefficient affected?
[Hint: See Example 54] [Ans. $r = 0.3$, it is not affe...]

(d) Variance-Covariance Method

This method of determining correlation coefficient is based on covariance. In this method, the following formula is used to obtain correlation coefficient:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\ \sqrt{Var(Y)}}$$

Or

$$r = \frac{Cov(X,Y)}{\sigma_x . \sigma_y} = \frac{\frac{\Sigma XY}{N} - \bar{X}\bar{Y}}{\sigma_x . \sigma_y}$$

Where, $Cov(X,Y) = \frac{\Sigma xy}{N} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N} = \frac{\Sigma XY}{N} - \bar{X}\bar{Y}$

The formula can also be written as:

$$r = \frac{\Sigma xy}{N . \sigma_x . \sigma_y} \quad \text{where, } x = X - \bar{X},\ y = Y - \bar{Y}$$

Example 22. For two series X and Y, Cov $(X, Y) = 15$, Var $(X) = 36$, Var $(Y) = 25$, calculate the coefficient of correlation.

Solution: Given $Cov(X,Y) = 15$, $Var(X) = 36$, $Var(Y) = 25$

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)}\ \sqrt{Var(Y)}} = \frac{15}{\sqrt{36}\ \sqrt{25}}$$

$$= \frac{15}{\sqrt{900}} = \frac{15}{30} = +0.50$$

Example 23. From the following data, compute the coefficient of correlation between X and Y:

X-series	*Y-series*
N = 30	N = 30
$\bar{X} = 40$	$\bar{Y} = 50$
$\sigma_x = 6$	$\sigma_y = 7$
$\Sigma xy = 360$	

(Where, x and y are deviations from their respective means)

Solution: We are given $N = 30, \bar{X} = 40, \bar{Y} = 50, \sigma_x = 6, \sigma_y = 7, \Sigma xy = 360$

Karl Pearson's coefficient of correlation is given by:

$$r = \frac{\Sigma xy}{N . \sigma_x . \sigma_y}$$

$$= \frac{360}{30 \times 6 \times 7} = \frac{360}{1260} = \frac{2}{7} = +0.286$$

Aliter: $r = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y}$, $Cov(X,Y) = \frac{\Sigma xy}{N} = \frac{360}{30} = 12$ where, $x = X - \bar{X}, y = Y - \bar{Y}$

$$r = \frac{12}{6 \times 7} = \frac{12}{42} = +0.286$$

IMPORTANT TYPICAL EXAMPLE

Example 24. *From two series X and Y, Cov (X,Y) = 25, r = 0.6, variance of X=36. Calculate standard deviation of y.*

Solution: Given, $Cov(X,Y) = 25, r = 0.6, var(X) = 36 \Rightarrow \sigma_x = \sqrt{36} = 6$. $[\because \sigma = \sqrt{\text{variance}}]$

$$r = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y}$$

$$+0.6 = \frac{25}{6 \times \sigma_y}$$

$$(0.6)(6 \times \sigma_y) = 25$$

$$(3.6)(\sigma_y) = 25$$

$$\sigma_y = \frac{25}{3.6} = 6.94$$

EXERCISE 1.6

1. The following results are obtained regarding two series. Compute coefficient of correlation

	X-series	Y-series
No. of items:	15	15
Arithmetic mean:	25	18
Standard deviation:	3.01	3.03

Sum of products of deviations of X and Y series from their means = 122. [Ans. r =

2. Calculate the coefficient of correlation where Cov (X,Y) = 488; Variance of X = 824 and Variance of Y = 325. [Ans. r =

3. If covariance between X and Y is 10 and the variance of X and Y are 16 and 9 respectively find the coefficient of correlation. [Ans. r =

4. Karl Pearson's coefficient of correlation between two variables X and Y is 0.64, the covariance is 16. If the variance of X is 9, find the standard deviation of Y-series. [Ans. σ_y =

5. The coefficient of correlation between two variables X and Y is 0.48 and their covariance 36. If the variance of X-series is 16, find the second moment about mean of Y-series. *[i.e., variance of Y-series].* [Ans. $\sigma_y^2 = \mu_2$ =

(B) Calculation of Coefficient of Correlation in Grouped Data/Bivariate Distribution

When number of items in two series is very large, then we present them by means of a two-way frequency table. This table gives the frequency distribution of two variables X and Y. The class intervals for Y-variables are presented in column heading (captions) and class intervals for X-variables are presented in row headings (stubs). Frequencies of the each cell of the table are counted by means of using tally bars.

Correlation coefficient in case of grouped data is computed by using the following formula:

$$r = \frac{\Sigma fdxdy - \frac{\Sigma fdx \cdot \Sigma fdy}{N}}{\sqrt{\Sigma fdx^2 - \frac{(\Sigma fdx)^2}{N}} \sqrt{\Sigma fdy^2 - \frac{(\Sigma fdy)^2}{N}}}$$

Or

$$r = \frac{N \times \Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2} \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

Steps

(1) Step deviations of X-variables are worked out and these are denoted by '*dx*'. Similarly, step deviations of Y-variables are calculated and these are denoted by '*dy*'.

(2) Step deviations of X-variables are multiplied by the corresponding frequencies and added up to get Σfdx. Similarly Σfdy is obtained.

(3) By multiplying the squared deviations of X-variables with the corresponding frequencies or multiplying Σfdx by *dx* and adding up, we get $\Sigma\, fdx^2$. Similarly $\Sigma\, fdy^2$ are obtained.

(4) Multiplying *dx* and *dy* and further multiplying them with their corresponding cell frequencies yields *fdxdy*. This product is written in the cell down at the right side/corner. Adding together all the cornered values vertically and horizontally gives $\Sigma fdxdy$.

(5) Putting the values of Σfdx, $\Sigma\, fdx^2$, $\Sigma\, fdy^2$ and $\Sigma fdxdy$ in the above formula to obtain correlation coefficient.

The following examples make clear the computation of correlation in grouped data:

Example 25. 30 pairs of X and Y are given below:

X:	14	20	33	25	41	18	24	29	38	45
Y:	147	242	296	312	518	196	214	340	492	568
X:	23	32	37	19	28	34	38	29	44	40
Y:	382	400	288	292	431	440	500	512	415	514
X:	22	39	43	44	12	27	39	38	17	26
Y:	382	481	516	598	122	200	451	387	245	413

Prepare a correlation table taking class interval of X as 10 to 20, 20 to 30, etc. and that of Y as 100 to 200, 200 to 300, etc. and find Karl Pearson's coefficient of correlation.

Solution:

Preparation of Bivariate Frequency Distribution

Y/X →	10—20	20—30	30—40	40—50	Total
100—200	|||(3)				3
200—300	|| (2)	||| (3)	||(2)		7
300—400		||||(4)	| (1)		5
400—500		|| (2)	⧻ (5)	| (1)	8
500—600		| (1)	|(1)	⧻ (5)	7
Total	5	10	9	6	N=30

(Lanscape Table Given at Page 35)

Applying the formula,

$$r = \frac{N \times \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2}\ \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

$$= \frac{30 \times 35 - (16)(9)}{\sqrt{30 \times 38 - (16)^2}\ \sqrt{30 \times 55 - (9)^2}}$$

$$= \frac{1050 - 144}{\sqrt{1140 - 256}\ \sqrt{1650 - 81}} = \frac{906}{\sqrt{884}\ \sqrt{1569}}$$

$$= \frac{906}{29.73 \times 39.61} = \frac{906}{1177.60} = 0.76$$

Example 26. Calculate Karl Pearson's coefficient of correlation from the following data:

X/Y	10—25	25—40	40—55
0-20	10	4	6
20-40	5	40	9
40-60	3	8	15

Solution: **(Landscape Table Given at Page 36)**

$$r = \frac{N \times \Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2}\ \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

$$= \frac{100 \times 16 - (6)(12)}{\sqrt{100 \times 46 - (6)^2}\ \sqrt{100 \times 48 - (12)^2}}$$

$$= \frac{1600 - 72}{\sqrt{4564}\ \sqrt{4656}}$$

$$= \frac{1528}{4609.77} = 0.33$$

Correlation Table of Solution 25

Let $dx = \frac{X - 25}{10}$, $dy = \frac{Y - 350}{100}$

Y ↓ M.V. \ X → M.V.	dx	−200	−100	0	+100	+200	f	fdy	fdy²	fdxdy
		−2	−1	0	+1	+2				
100-200	150	2 3 6	1 2 2	—	—	—	5	−5	5	8
200-300	250	0 3 0	−1 2 −2	—	—	—	10	0	0	0
300-400	350	0 4 0	0 1 0	2 1 2	1 5 5	4 5 20	6	12	24	22
400-500	450	0 2 0	1 5 5	2 1 2	1 5 5	4 5 20	9	9	9	5
500-600	550	0 1 0	2 1 2	2 1 2	1 5 5	4 5 20	9	9	9	5
							N =30	Σ fdx =16	Σ fdx² =38	Σ fdxdy =35
								Σ fdy =9	Σ fdy² =55	Σ fdxdy =35

Let $dx = \frac{X - 30}{20}$, $dy = \frac{Y - 32.50}{15}$

Correlation Table of Solution 26

X ↓ M.V.	dx	Y → M.V.	dy	10—25 (17.5, −15, −1)	25—40 (32.50, 0, 0)	40—55 (47.5, +15, +1)	f	fdx	fdx²	fdxdy
0-20	10	−20	−1	1 / 10 / 10	0 / 4 / 0	−1 / 6 / −6	20	−20	20	4
20-40	30	0	0	0 / 5 / 0	0 / 40 / 0	0 / 9 / 0	54	0	0	0
40-60	50	+20	+1	−1 / 3 / −3	0 / 8 / 0	1 / 15 / 15	26	26	26	12
			f	18	52	30	N = 100	Σ fdx =6	Σ fdx² =46	Σ fdxdy =16
			fdy	−18	0	30	Σ fdy =12			
			fdy²	18	0	30	Σ fdy² =48			
			fdxdy	7	0	9	Σ fdxdy =16			

Let $dx = X - 20, dy = \frac{Y - 12.5}{5}$

Correlation Table of Solution 27

Y ↓ M.V.	dy	X → M.V.	dx	18 (−2)	19 (−1)	20 (0)	21 (+1)	22 (+2)	f	fdy	fdy²	fdxdy
0-5	2.5	−10	−2	—	—	—	−2 / 3 / −6	−4 / 1 / −4	4	−8	16	−10
5-10	7.5	−5	−1	—	—	—	−1 / 3 / −3	−2 / 2 / −4	5	−5	5	−7
10-15	12.5	0	0	—	—	0 / 7 / 0	0 / 10 / 0	—	17	0	0	0
15-20	17.5	+5	+1	—	−1 / 5 / −5	0 / 4 / 0	—	—	9	9	9	−5
20-25	22.5	+10	+2	−4 / 3 / −12	−2 / 2 / −4	—	—	—	5	10	20	−16
			f	3	7	11	16	3	N = 40	Σ fdy = 6	Σ fdy² = 50	Σ fdxdy = −38
			fdx	−6	−7	0	16	6	Σ fdx = 9			
			fdx²	12	7	0	16	12	Σ fdx² = 47			
			fdxdy	−12	−9	0	−9	−8	Σ fdxdy = −38			

Example 27. Calculate Karl Pearson's coefficient of correlation from the following data:

Y/X	18	19	20	21	22	Total
0—5	—	—	—	3	1	4
5—10	—	—	—	3	2	5
10—15	—	—	7	10	—	17
15—20	—	5	4	—	—	9
20—25	3	2	—	—	—	5
Total	3	7	11	16	3	40

Solution: **(Landscape Table Given at Page 37)**

$$r = \frac{N \times \Sigma fdxdy - (\Sigma fdx)(\Sigma fdy)}{\sqrt{N \times \Sigma fdx^2 - (\Sigma fdx)^2}\ \sqrt{N \times \Sigma fdy^2 - (\Sigma fdy)^2}}$$

$$= \frac{40(-38) - (6)(9)}{\sqrt{40(47) - (9)^2}\ \sqrt{40(50) - (6)^2}}$$

$$= \frac{-1574}{\sqrt{1799}\ \sqrt{1964}} = \frac{-1574}{42.41 \times 44.32} = \frac{-1574}{1879.61}$$

$= -0.837'0 \approx -0.84.$

It shows a high degree of negative correlation between X and Y.

EXERCISE 1.7

1. Calculate Karl Pearson's coefficient of correlation for the following distribution:

Y \ X	200—300	300—400	400—500	500—600	600—700
10—15	–	–	–	3	7
15—20	–	4	9	4	3
20—25	7	6	12	5	–
25—30	3	10	19	8	–

Also calculate its probable error. [Ans. $r = -0.438$, PE = 0.0...

2. Calculate the coefficient of correlation between marks and age from the following data:

Marks \ Age	18	19	20	21
200—250	4	4	2	1
250—300	3	5	4	2
300—350	2	6	8	5
350—400	1	4	6	10

Can we conclude that increase in age causes increase in marks? [Ans. r = 0.8...

3. 24 pairs of X and Y are given below:

X	15	0	1	3	16	2	18	5	
Y	13	1	2	7	8	9	12	9	
X	4	17	6	19	14	9	8	13	
Y	17	16	6	18	11	3	5	4	
X	10	13	11	11	12	18	9	7	
Y	10	11	14	7	18	15	15	3	

Prepare a correlation table taking the magnitude of each class interval as four and the first interval as equal to 0 and less than 4. Calculate Karl Pearson's coefficient between X and Y. [Ans. $r = 0.578$]

4. The frequency distribution of marks obtained in Physics and Chemistry by 100 students are given in the following table. Determine:

(*i*) Percentage of students passed in Physics and Chemistry, while for passing minimum 60% is required.

(*ii*) Coefficient of correlation.

Physics \ Chemistry	40—49	50—59	60—69	70—79	80—89	90—99	Total
90—99	—	—	—	2	4	4	10
80—89			1	4	6	5	16
70—79	—	—	5	10	8	1	24
60—69	1	4	9	5	2	—	21
50—59	3	6	6	2	—	—	17
40—49	3	5	4	—	—	—	12
Total	7	15	25	23	20	10	100

[Ans. (*i*) % of students in Physics = 71%, % of students passed in Chemistry = 78%, (*ii*) $r = 0.8056$]

Assumptions of Karl Pearson' Coefficient of Correlation

Karl Pearson's coefficient of correlation is based on the following assumptions:

(1) Affected by a Large Number of Independent Causes: Series or variables which are correlated, are affected by a large number of factors that result in a normal distribution.

(2) Cause and Effect Relation: There is a cause and effect relationship between the forces affecting the distribution of the items in the two series.

(3) Linear Relationship: Two variables are linearly related. Plotting the values of the variables in a scatter diagram yields a straight line.

Properties of the Coefficient of Correlation

The following are the important properties of the correlation coefficient (r):

(1) **Limits of Coefficient of Correlation:** Karl Pearson's coefficient of correlation lies between −1 and +1. Symbolically

$$-1 \le r \le +1$$

This implies r can never exceed +1 and never becomes less than −1. It always lies between −1 and +1.

(2) **Change of Origin and Scale:** Shifting the origin or scale does not affect in any way the value of correlation coefficient. coefficient of correlation is independent of the change of origin and scale. If the scale of a series is changed or the origin is shifted, then correlation coefficient remains unchanged.

(3) **Geometric Mean of Regression Coefficients:** Correlation coefficient is the geometric mean of the regression coefficients byx and bxy. Symbolically:

$$r = \sqrt{bxy \cdot byx}$$

(4) If X and Y are independent variables, then coefficient of correlation is zero but the converse is not necessarily true. [For proof, See Example 55].

(5) **Pure Number:** 'r' is a pure number and is independent of the units of measurements. This implies that even if the two variables are expressed in two different units of measurements viz. rainfall in inches, and yields of crops in quintals, the value of correlation coefficient comes out with a pure number. Thus, it does not require that the units of both the variables should be the same.

(6) **Symmetric:** The coefficient of correlation between the two variables x and y is symmetric, i.e., $r_{xy} = r_{yx}$. It means that either we compute the value of correlation coefficient between x and y or between y and x, the coefficient of correlation remains the same.

Interpreting the Coefficient of Correlation

Coefficient of correlation measures the degree of relationship between two variables. It is denoted by 'r'. The value of correlation coefficient lies between −1 and +1. The value of correlation coefficient can be interpreted in the following ways:

(i) If $r = +1$, then there is perfect positive correlation.
(ii) If $r = 0$, then there is absence of linear correlation.
(iii) If $r = +0.25$, then there will be low degree of positive correlation.
(iv) If $r = +0.50$, then there is moderate degree of correlation.
(v) If $r = +0.75$, then there is high degree of positive correlation.

Similarly, negative values of r can be interpreted.

Probable Error and Karl Pearson's Coefficient of Correlation

To test the reliability of Karl Pearson's correlation coefficient, probable error is used. The following formula is used to determine probable error:

$$\text{Probable Error (P.E.)} = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

Where, r is the coefficient of correlation and N, the number of pairs of observations.

If the constant 0.6745 is omitted from the above formula of probable error, we get the **standard error of the coefficient of correlation**. Thus,

$$SE_r = \frac{1 - r^2}{\sqrt{N}}$$

Utility of Probable Error: (1) Probable error is used to interpret the value of the correlation coefficient. Interpretation of r with the help of probable error is made clear by the following points:

(i) If $|r| > 6$ P.E., then coefficient of correlation (r) is taken to be significant.
(ii) If $|r| < 6$ P.E., then coefficient of correlation (r) is taken to be insignificant. This means that, there is no evidence of the existence of correlation in both the series.

(2) Probable error also determines the upper and lower limits within which the correlation of a randomly selected sample from the same universe will fall. Symbolically,

Upper Limit = r + P.E.$_r$, Lower Limit = r − P.E.

Example 28. Find the Karl Pearson's coefficient of correlation from the following data:

X:	9	28	45	60	70	50
Y:	100	60	50	40	33	57

Also calculate probable error and point out whether the coefficient of correlation is significant or not.

Solution:

Calculation of Coefficient of Correlation

X	dx	dx^2	Y	dy	dy^2	$dxdy$
9	−36	1296	100	50	2500	−1800
28	−17	289	60	10	100	−170
45 = A	0	0	50 = A	0	0	0
60	15	225	40	−10	100	−150
70	25	625	33	−17	289	−425
50	5	25	57	7	49	35
$N = 6$	$\Sigma dx = -8$	$\Sigma dx^2 = 2460$		$\Sigma dy = 40$	$\Sigma dy^2 = 3038$	$\Sigma dxdy = -2510$

$$r = \frac{N \times \Sigma dxdy - \Sigma dx \cdot \Sigma dy}{\sqrt{N \times \Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N \times \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{6 \times (-2510) - (-8)(40)}{\sqrt{6 \times 2460 - (-8)^2}\ \sqrt{6 \times 3038 - (40)^2}}$$

$$= \frac{-15060+320}{\sqrt{14760-64}\ \sqrt{18228-1600}} = \frac{-14740}{\sqrt{14696}\ \sqrt{16628}}$$

$$= \frac{-14740}{121.227 \times 128.95} = \frac{-14740}{15632.221} = -0.94$$

Calculation of P.E.

$$\text{P.E.} = 0.6745 \times \frac{1-r^2}{\sqrt{N}} = 0.6745 \times \frac{1-(-0.94)^2}{\sqrt{6}}$$

$$= 0.6745 \times \frac{0.1164}{2.449} = 0.03205$$

Significance of r

$$\frac{|r|}{P.E.} = \frac{0.94}{0.03205} = 29.32$$

$$\Rightarrow \quad |r| = 29.32 \text{ P.E.}$$

Since, $|r|$ is more than 6 times the P.E., so, correlation coefficient is highly significant.

Example 29. A student calculates the value of r as 0.7 when the value of n is 5 and concludes that r is highly significant. Is he correct?

Solution: We know that if the value of $r > 6$ P.E., then it is considered to be significant.

$$\text{P.E.} = 0.6745 = \frac{1-r^2}{\sqrt{N}}$$

$$= 0.6745 \times \frac{1-(0.7)^2}{\sqrt{5}} = 0.15$$

Now, $$\frac{r}{P.E.} = \frac{0.7}{0.15} = 4.67 \quad \Rightarrow \quad r = 4.67\ P.E.$$

Since r is less than six times the P.E., r is insignificant and the student is wrong in his calculation.

Example 30. Show by calculation which 'r' is more significant: (*i*) $r = 0.90$, P.E. = 0.03 (*ii*) $r = 0.70$, P.E. = 0.02.

Solution: r is most significant in that case in which it is the highest number of times the P.E. is compared as below:

(*i*) $\frac{r}{P.E.} = \frac{0.90}{0.03} = 30$, so r is 30 times of P.E.

(*ii*) $\frac{r}{P.E.} = \frac{0.70}{0.02} = 35$, so r is 35 times of P.E.

It is clear from the above that coefficient of correlation is the most significant in case (ii).

EXERCISE 1.8

1. Find Karl Pearson's Coefficient of correlation from the following series of marks secured by 10 students in a class test in Mathematics and Statistics.

Maths (X):	45	70	65	30	90	40	50	75	85	60
Statistics (Y):	35	90	70	40	95	40	60	80	80	50

Also calculate probable error. Is the value of r significant or not?
[Ans. $r = 0.903$, P.E. = 0.039, Highly significant]

2. Calculate the coefficient of correlation between the heights of fathers and sons from the following:

Height of Fathers (inches):	65	66	67	68	69	70	71
Height of Sons (inches):	67	68	66	69	72	72	69

Also calculate its probable error. Is the value of r significant or not?
[Ans. $r = 0.668$, P.E. = 0.141, Not significant]

3. (a) Find r if $N = 100$, P.E. = 0.05 (b) Find N if P.E. = 0.025, $r = .80$ [Ans. (*a*) $r = 0.5086$ (b) $N = 94$]
4. Comment on the significance of r in the following situations:
   (*i*) $N = 25$, $r = 0.8$
   (*ii*) $N = 100$, P.E. = 0.04 [Ans. (*i*) P.E. = 0.049, significant (*ii*) $r = 0.63$, significant]
5. The correlation coefficient of a sample of 100 pairs of items was 0.92. Within what limits does it hold good for another sample taken from the same universe?
[Ans. PE = 0.0103, 0.92 ± 0.0103]

(*ii*) Spearman's Rank Correlation Method

This method of determining correlation was propounded by Prof. Spearman in 1904. By this method, correlation between qualitative data namely beauty, honesty, intelligence, etc., can be computed. Such types of variables can be assigned ranks but their quantitative measurement is not possible. Thus, rank correlation method is used in such cases. The following is the formula for the computation of rank correlation coefficient:

$$R = 1 - \frac{6\Sigma D^2}{N(N^2-1)} \quad \text{or} \quad 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Where, R = Rank coefficient of correlation, D = Difference between two ranks $(R_1 - R_2)$,
N = Number of pair of observations.

The value of rank correlation coefficient always lies between –1 and +1.

Note: 1. The value of rank correlation coefficient will be equal to the value of Pearson's Coefficient of Correlation for the two characteristics taking the ranks as values of the variables, provided no rank value is repeated *i.e.* the rank values of all the variables are different.

2. The sum total of rank difference (*i.e.*, ΣD) is always equal to zero,
i.e., $\Sigma D = \Sigma(R_1 - R_2) = 0$. This serves as check on the calculation work.

This method can be studied in the following three different situations:

(1) When ranks are given

(2) When ranks are not given

(3) When equal or tied ranks.

▶ **(1) When ranks are given**

When ranks are given, the following procedure is adopted to find the rank correlation coefficient:

(*i*) Ranks difference is found out by deducting the ranks of Y series from the corresponding ranks of X series. This is denoted by D, *i.e.*, $D = R_1 - R_2$.

(*ii*) Squaring the rank differences and summing them up, we get ΣD^2.

(*iii*) Finally, the following formula is used:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

The following examples make the above said method clear:

Example 31. In a fancy-dress competition, two judges accorded the following ranks to eight participants:

Judge X:	8	7	6	3	2	1	5	4
Judge Y:	7	5	4	1	3	2	6	8

Calculate coefficient of rank correlation.

Solution:

Calculation of Rank Correlation Coefficient

Judge X R_1	Judge Y R_2	$D = R_1 - R_2$	D^2
8	7	+1	1
7	5	+2	4
6	4	+2	4
3	1	+2	4
2	3	−1	1
1	2	−1	1
5	6	−1	1
4	8	−4	16
$N = 8$		$\Sigma D = 0$	$\Sigma D^2 = 32$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 32}{8^3 - 8} = 1 - \frac{192}{504}$$

$$= 1 - 0.381 = 0.619$$

There is, thus, moderate degree of positive relationship between the two judgements.

Example 32. Two ladies were asked to rank 10 different types of lipsticks. The ranks given by them are given below:

Lipsticks:	A	B	C	D	E	F	G	H	I	J
Neelu:	1	6	3	9	5	2	7	10	8	4
Neena:	6	8	3	7	2	1	5	9	4	10

Calculate Spearman's rank correlation coefficient.

Solution:

Calculation of Rank Correlation Coefficient

R_1	R_2	$D = R_1 - R_2$	D^2
1	6	−5	25
6	8	−2	4
3	3	0	0
9	7	+2	4
5	2	+3	9
2	1	+1	1
7	5	+2	4
10	9	+1	1
8	4	+4	16
4	10	−6	36
$N = 10$		$\Sigma D = 0$	$\Sigma D^2 = 100$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 100}{10^3 - 10} = 1 - \frac{600}{990}$$

$$= 1 - \frac{60}{99} = 1 - 0.606 = 0.394$$

Example 33. Ten competitors in a beauty contest are ranked by three judges in the following order:

1st Judge	1	6	5	10	3	2	4	9	7	8
2nd Judge	3	5	8	4	7	10	2	1	6	9
3rd Judge	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient to determine which pair of judges has the nearest approach to common tastes in beauty.

Solution: In order to find out which pair of judges has the nearest approach to common tastes in beauty, we compare the rank correlation coefficient between the judgements of

(*i*) 1st Judge and 2nd Judge

(*ii*) 2nd Judge and 3rd Judge

(*iii*) 1st Judge and 3rd Judge.

Calculation of Rank Correlation Coefficient

Rank by 1st Judge (R_1)	Rank by 2nd Judge (R_2)	Rank by 3rd Judge (R_3)	$(R_1 - R_2)^2$ D_{12}^2	$(R_2 - R_3)^2$ D_{23}^2	$(R_1 - R_3)^2$ D_{13}^2
1	3	6	4	9	25
6	5	4	1	1	4
5	8	9	9	1	16
10	4	8	36	16	4
3	7	1	16	36	4
2	10	2	64	64	1
4	2	3	4	1	1
9	1	10	64	81	4
7	6	5	1	1	1
8	9	7	1	4	$\Sigma D_{13}^2 = 86$
$N = 10$	$N = 10$	$N = 10$	$\Sigma D_{12}^2 = 200$	$\Sigma D_{23}^2 = 214$	

Applying the formula,

$$R_{12} = 1 - \frac{6\Sigma D_{12}^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 200}{10^3 - 10} = 1 - \frac{1200}{990} = -0.212$$

$$R_{23} = 1 - \frac{6\Sigma D_{23}^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 214}{10^3 - 10} = 1 - \frac{1284}{990} = -0.297$$

$$R_{13} = = 1 - \frac{6\Sigma D_{13}^2}{N^3 - N}$$

$$= 1 - \frac{6 \times 60}{10^3 - 10} = 1 - \frac{360}{990} = +0.636$$

Since the coefficient of rank correlation is positive and maximum in the judgement of the first and third judges, we conclude that they have the nearest approach to common tastes in beauty.

▶ **(2) When ranks are not given**

When we are given the actual data and not the ranks, the following procedure is adopted to find out rank correlation coefficient:

(i) First of all, ranks are assigned to the items of X and Y series on the basis of their size. The largest value is assigned rank first, second largest second rank and similarly other values

are ranked. Sometimes, the smallest value is assigned the highest rank *i.e.* in descending order of the values. However, the same order (*i.e.* ascending order or descending order) of assigning the ranks must be maintained in both the series.

(ii) Rank difference of both the series ($D = R_1 - R_2$) is found and squared up. The squared rank difference, thus obtained is summed upto get ΣD^2.

(iii) Finally, the following formula is used to obtain rank correlation coefficient:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

The following example gives clarity to the above said method and its procedure.

Example 14. Find out the coefficient of correlation between X and Y by the method of rank differences:

X:	15	17	14	13	11	12	16	18	10	9
Y:	18	12	4	6	7	9	3	10	2	5

Solution:

Calculation of Rank Correlation Coefficient

X	Rank R_1	Y	Rank R_2	$D = R_1 - R_2$	D^2
15	4	18	1	+3	9
17	2	12	2	0	0
14	5	4	8	−3	9
13	6	6	6	0	0
11	8	7	5	+3	9
12	7	9	4	+3	9
16	3	3	9	−6	36
18	1	10	3	−2	4
10	9	2	10	−1	1
9	10	5	7	+3	9
$N = 10$				$\Sigma D = 0$	$\Sigma D^2 = 86$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

Here, $N = 10$, $\Sigma D^2 = 86$

$$R = 1 - \frac{6 \times 86}{10^3 - 10}$$

$$= 1 - \frac{516}{990} = 1 - 0.52 = 0.48$$

Thus, there is positive correlation between X and Y.

(3) When equal or tied ranks

When two or more items have equal values in a series, then in such case, items of equal value are assigned common ranks, which is average of the ranks. For example, when item 10 appears twice in a series and their rank turns out to be 7 and 8 respectively, then they should be assigned $\frac{7+8}{2} = 7.5$ rank. In such case, some modification has to be made in the formula. Here, the following formula is used to determine rank correlation coefficient:

$$R = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \ldots\ldots]}{N^3 - N}$$

Here, m = Number of items of equal ranks.

The correction factor of $\frac{1}{12}(m^3 - m)$ is added to ΣD^2 for such number of times as the cases of equal ranks in the question.

Example 35. Calculate coefficient of rank correlation from the following data:

X	15	10	20	28	12	10	16	18
Y	16	14	10	12	11	15	18	13

Solution:

Make corrections for tied ranks.

Calculation of Coefficient of Rank Correlation

X	R_1	Y	R_2	$D = R_1 - R_2$	D^2
15	5	16	2	3	9.00
10	7.5	14	4	3.5	12.25
20	2	10	8	−6	36.00
28	1	12	5.5	−4.5	20.25
12	6	11	7	−1	1.00
10	7.5	15	3	4.5	20.25
16	4	18	1	3	9.00
18	3	12	5.5	−2.5	6.25
$N = 8$				$\Sigma D = 0$	$\Sigma D^2 = 114$

In this question, the cases of equal rank are two, one for X series and other for Y series. Hence $\frac{1}{12}(m^3 - m)$ would be added for two times in ΣD^2.

Here, number 10 is repeated twice in series X and number 12 is repeated twice in series Y. Therefore, in both X and Y, $m = 2$.

$$R = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m)]}{N^3 - N}$$

$$= 1 - \frac{6[114 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)]}{8^3 - 8}$$

$$= 1 - \frac{6[114 + \frac{1}{12}(6) + \frac{1}{12}(6)]}{512 - 8} = 1 - \frac{6[114 + 0.5 + 0.5]}{504}$$

$$= 1 - \frac{6[115]}{504} = 1 - \frac{690}{504} = 1 - 1.369 = -0.369$$

Example 36. Calculate coefficient of correlation by means of ranking method from the following data:

X:	40	50	60	60	80	50	70	60
Y:	80	120	160	170	130	200	210	130

Solution:

Calculation of Rank Coefficient of Correlation

X	R_1	Y	R_2	$D = R_1 - R_2$	D^2
40	8	80	8	0	0
50	6.5	120	7	−0.5	0.25
60	4	160	4	0	0
60	4	170	3	1	1
80	1	130	5.5	−4.5	20.25
50	6.5	200	2	4.5	20.25
70	2	210	1	1	1
60	4	130	5.5	−1.5	2.25
$N = 8$				$\Sigma D = 0$	$\Sigma D^2 = 45.00$

In this question in X series, the values 60 and 50 are repeated thrice and twice. The average rank for the value 60 is 4 (3 + 4 + 5 ÷ 3) while for the value 50 it is 6.5 (6 + 7 ÷ 2). In both the cases, the correlation factor will be $\frac{1}{12}(3^3 - 3)$ and $\frac{1}{12}(2^3 - 2)$. In series Y, the 130 is repeated twice. The average rank for the value 130 is 5.5 (5 + 6 ÷ 2). In this case, correction factor will be $\frac{1}{12}(2^3 - 2)$.

Applying the formula,

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

$\Sigma D^2 = 45$, $m_1 = 3$, $m_2 = 2$, $m_3 = 2$, $N = 8$

By subsituting values in the above formula, we get

$$R = 1 - \frac{6\left[45 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]}{8(8^2 - 1)}$$

$$= 1 - \frac{6(45 + 2 + 0.5 + 0.5)}{8(63)} = 1 - \frac{6(48)}{504} = 1 - \frac{288}{504}$$

$$= 1 - 0.571 = 0.429$$

IMPORTANT TYPICAL EXAMPLES

Example 37. The ranks of the same 8 students in tests in Mathematics and Statistics were as follow the two numbers within brackets denoting the ranks of the same students Mathematics and Statistics respectively:

(1, 4), (2, 2), (3, 1), (4, 6), (5, 8), (6, 3), (7, 5), (8, 7)

(*i*) Calculate the rank correlation for proficiencies of this group in Math's and Statistics

(*ii*) What does the value of the coefficient obtained indicates?

(*iii*) If you have found out Karl Pearson's simple coefficient of correlation between the ranks of these 16 students. Would your results have been the same as obtained in (*i*) or any different?

Solution: (*i*)

Calculation of Rank Correlation Coefficient

Ranks in Maths (R_1)	Ranks in Statistics (R_2)	$D = R_1 - R_2$	D^2
1	4	−3	9
2	2	0	0
3	1	+2	4
4	6	−2	4
5	8	−3	9
6	3	+3	9
7	5	+2	4
8	7	+1	1
$N = 8$			$\Sigma D^2 = 40$

Applying the formula

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 40}{8^3 - 8} = 1 - \frac{240}{504}$$

$$= \frac{504 - 240}{504} = \frac{264}{504} = +0.523$$

(*ii*) The value of rank correlation coefficient indicates that there is moderate degree of positive correlation.

(*iii*)

Calculation of Karl Pearson's Coefficient of Correlation

Ranks in Maths (X)	A = 4 dx	dx^2	Ranks in Statistics (Y)	A = 4 dy	dy^2	$dx\,dy$
1	−3	9	4	0	0	0
2	−2	4	2	−2	4	4
3	−1	1	1	−3	9	3
4 = A	0	0	6	+2	4	0
5	+1	1	8	+4	16	4
6	+2	4	3	−1	1	−2
7	+3	9	5	+1	1	3
8	+4	16	7	+3	9	12
$N = 8$	$\Sigma dx = 4$	$\Sigma dx^2 = 44$		$\Sigma dy = 4$	$\Sigma dy^2 = 44$	$\Sigma dxdy = 24$

Applying the formula

$$r = \frac{N \cdot \Sigma dxdy - \Sigma dx \cdot \Sigma dy}{\sqrt{N \cdot \Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N \cdot \Sigma dy^2 - (\Sigma dy)^2}}$$

$$= \frac{8 \times 24 - (4)(4)}{\sqrt{8 \times 44 - (4)^2}\ \sqrt{8 \times 44 - (4)^2}}$$

$$= \frac{192 - 16}{\sqrt{336}\ \sqrt{336}} = \frac{176}{336} = 0.523$$

It is evident that the value of correlation coefficient computed by using Karl Pearson is the same as obtained by rank correlation method. The reason is that when the ranks of the students are not repeated, then the two methods give the same answer.

Example 38. Calculate rank correlation coefficient from the following data:

Serial No.:	1	2	3	4	5	6	7	8	9	10
Rank Difference:	−2	?	−1	+3	+2	0	−1	+3	+3	−2

Solution: The total of rank differences (ΣD) is always equal to zero and on this base the missing rank difference will be calculated. Let the missing item be 'a'.

As $\Sigma D = 0 \Rightarrow -2 + a - 1 + 3 + 2 + 0 - 4 + 3 + 3 - 2 = 0$

$\therefore \quad a = -2$

Calculation of Coefficient of Rank Correlation

Sr. No.	Rank Difference D	D^2
1	–2	4
2	–2	4
3	–1	1
4	+3	9
5	+2	4
6	0	0
7	–4	16
8	+3	9
9	+3	9
10	–2	4
$N = 10$	$\Sigma D = 0$	$\Sigma D^2 = 60$

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 60}{10^3 - 10}$$

$$= 1 - \frac{360}{990} = 1 - 0.364 = +0.636$$

Example 39. The coefficient of rank correlation of marks obtained by 10 students in English and Mathematics was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Solution: Given, R = 0.5, $N = 10$, Incorrect difference of ranks (D) = 3

Correct difference of ranks (D) = 7

We know that:

$$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$0.5 = 1 - \frac{6\Sigma D^2}{10^3 - 10}$$

$$0.5 = 1 - \frac{6\Sigma D^2}{990}$$

$$\frac{6\Sigma D^2}{990} = 1 - 0.5 = 0.5$$

$\Rightarrow$ Incorrected $\Sigma D^2 = 82.5$

Corrected ΣD^2 = 82.5 – (Incorrect value)^2 + (Correct value)^2

$= 82.5 - 3^2 + 7^2 = 122.5$

Corrected Coefficient of Rank Correlation (R) $= 1 - \frac{6\Sigma D^2}{N^3 - N}$

$$= 1 - \frac{6 \times 122.5}{10^3 - 10} = 1 - \frac{735}{990} = 0.258$$

Thus, the correct value of rank correlation coefficient is 0.258.

Example 40. The rank correlation coefficient between marks obtained by some students in 'Statistics' and 'Accountancy' is found to be 0.8. If the total of squares of rank differences is 33, find the number of students.

Solution: Given, $R = 0.8$, $\Sigma D^2 = 33$

Now, $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$0.8 = 1 - \frac{6 \times 33}{N^3 - N}$$

$$\frac{198}{N^3 - N} = 1 - 0.8 = 0.2$$

$$\Rightarrow \quad N^3 - N = \frac{198}{0.2} = 990$$

$$N(N^2 - 1) = 990 \qquad [\because\ a^2 - b^2 = (a+b)(a-b)]$$

$$N(N+1)(N-1) = 990$$

$$(N-1)(N)(N+1) = 9 \times 10 \times 11 \qquad \text{Comparing both sides, we get:}$$

$$\therefore \quad N - 1 = 9$$

$$\Rightarrow \quad N = 10$$

Example 41. The rank correlation coefficient between marks obtained by 10 students in Mathematics and Economics was found to be 0.5. Find the sum of squares of differences of ranks.

Solution: Given, $R = 0.5$, $N = 10$

Now, $$R = 1 - \frac{6\Sigma D^2}{N^3 - N}$$

$$\frac{6\Sigma D^2}{N^3 - N} = 1 - R$$

$$\frac{6\Sigma D^2}{10^3 - 10} = 1 - 0.5 = 0.5$$

$$6\Sigma D^2 = 0.5 \times 990$$

$$\Rightarrow \quad \Sigma D^2 = \frac{0.5 \times 990}{6} = 82.5$$

▶ Merits and Demerits of Rank Correlation Method

Merits

(1) This method is simple to understand and easy to apply as compared to Karl Pearson's method.

(2) When the data are of qualitative nature like beauty, honesty, intelligence, etc., this is the only method to be employed.

(3) When we are given the ranks and not the actual data, this method can be usefully employed.

Demerits

(1) This method cannot be used for finding correlation in a grouped frequency distribution.

(2) When the number of items exceed 30, the calculations become quite tedious and require a lot of time.

EXERCISE 1.9

1. Ten commerce graduates appeared before a selection board consisting of two members X and Y for the post of probationary officer in a certain bank. If the rank order of each of the members is given below, find out the coefficient of rank correlation:

Rank order by X:	1	6	5	10	3	2	4	9	7	8
Rank order by Y:	3	5	8	4	7	10	2	1	6	9

[Ans. $R = -0.21$]

2. Ten competitors in an intelligence test are ranked by three judges in the following order:

Judge I:	9	3	7	5	1	6	2	4	10	8
Judge II:	9	1	10	4	3	8	5	2	7	6
Judge III:	6	3	8	7	2	4	1	5	9	10

Use the rank correlation coefficient to determine:

(*i*) Which pair of judges agree the most?

(*ii*) Which pair of judges disagree the most?

[Ans. $R_{12} = 0.71$, $R_{23} = 0.467$, $R_{13} = 0.6$; (*i*) Ist and IIIrd (*ii*) IInd and IIIrd]

3. Find out the coefficient of correlation between X and Y by the method of rank differences:

X:	75	88	95	70	60	80	81	50
Y:	120	130	130	115	110	140	142	100

[Ans. R = 0.393]

4. Find out the coefficient of correlation between X and Y by the method of rank differences:

X:	46	56	39	45	54	58	36	49
Y:	30	60	40	50	70	70	30	50

[Ans. R = 0.7

5. Find the rank correlation coefficient from the following marks awarded by the examiners in statistics:

R. Nos.:	1	2	3	4	5	6	7	8	9	10	11
Marks Awarded by Examiner A:	24	29	19	14	30	19	27	30	20	28	11
Marks Awarded by Examiner B:	37	35	16	26	23	27	19	20	16	11	21
Marks Awarded by Examiner C:	30	28	20	25	25	30	20	24	22	29	15

[Ans. $R_{AB} = -0.027$, $R_{BC} = 0.5272$, $R_{AC} = 0.26136$]

6. From the following data, calculate Spearman's coefficient of correlation:

X:	80	78	75	75	68	67	60	59
Y:	12	13	14	14	14	16	15	17

[Ans. R = – 0.928]

7. The ranks of the same 16 students in tests in Mathematics and Statistics were as follows, the two numbers within brackets denoting the ranks of the same students in Mathematics and Statistics respectively:
(1, 1), (2, 10), (3, 3), (4, 4), (5, 5), (6, 7), (7, 2), (8, 6), (9, 8), (10, 11), (11, 15), (12, 9), (13, 14), (14, 12), (15, 16), (16, 13).

(*i*) Calculate the rank correlation for proficiencies of this group of Math's and Statistics.

(*ii*) What does the value of the coefficient obtained indicates?

(*iii*) If you have found out Karl Pearson's simple coefficient of correlation between the ranks of these 16 students would your results have been the same as obtained in (a) or any difference?

[Ans. R = 0.8, $r = 0.8$]

8. From the following data, calculate Spearman's coefficient of correlation:

Sr. No.:	1	2	3	4	5	6	7	8	9	10
Rank differences:	–2	–4	–1	+3	+2	0	?	+3	?3	–2

[Ans. R = +0.636]

9. The coefficient of rank correlation between debenture prices and share prices is found to be 0.143. If the sum of squares of the difference in ranks is given to be 48, find the value of N.

[Ans. N = 7]

(iii) CONCURRENT DEVIATION METHOD

Concurrent deviation method of determining the correlation is extremely simple method. In this method, correlation is determined on the basis of direction of the deviations. Under this method, taking into consideration the direction of deviations, they are assigned (+) or (–) or (0) signs. The following steps are taken to find out correlation in this method:

(1) Under this method, whatever the series X and Y are to be studied for correlation, each item of the series is compared with its preceding item. If the value is more than its preceding value, then its deviation is assigned (+) sign, if less than preceding value then (–) sign and if equal to the

preceeding value then (0) sign is assigned. After this, third item is compared with the second, fo item is compared with the third and this process goes on till the deviations of all items in a series worked out.

(2) The deviations of X and Y series (dx) and (dy) are multiplied to get dxdy. Product of t signs will be positive (+) and opposite signs will be negative (–) like:

(+) (+) = +,
(–) (–) = +,
(0) (0) = +,
(–) (+) = –,
(+) (–) = –,
(0) (–) = –,
(–) (0) = –,
(0) (+) = –

(3) Summing the positive dxdy signs, their number is counted. This is known as the numbe concurrent deviations. It is denoted by the sign 'C'. The deviations with minus signs are exclud from the computation. They are ignored. If all the deviations in a series have minus signs, t number of concurrent deviations will be zero *i.e.* C=0.

(4) Finally, the following formula is used for determining coefficient of concurrent deviati

$$r_c = \pm\sqrt{\pm\left(\frac{2C-n}{n}\right)}$$

Here, r_c =Coefficient of concurrent deviations;

C = Number of concurrent deviations or Number of positive signs obtained b multiplying *dx* with *dy*;

*n = Number of pairs of observations minus one = $N-1$.

Note: In this formula ± sign is used both inside and outside the radical sign. If the value of (2C– is positive, then (+) sign will be used both inside and outside the radical sign becaus such case correlation will be positive. On the contrary, if (2C–n) has negative sign, th minus sign will be used both inside and outside the radical sign because correlation will negative.

The value of coefficient of concurrent deviation always lies between –1 and +1.

The following examples make the procedure of concurrent deviation method clear.

Example 42. Find coefficient of concurrent deviation from the following data:

X:	85	91	56	72	95	76	89	51	59	9
Y:	18.3	20.8	16.9	15.7	19.2	18.1	17.5	14.9	18.9	15

** Since there is no sign for the first value of X and Y, n is always taken to be one less than the actu number of observations.*

Solution:

X	Deviation signs (*dx*)	Y	Deviation signs (*dy*)	*dxdy*
85		18.3		
91	+	20.8	+	+
56	–	16.9	–	+
72	+	15.7	–	–
95	+	19.2	+	+
76	–	18.1	–	+
89	+	17.5	–	–
51	–	14.9	–	+
59	+	18.9	+	+
90	+	15.4	–	–
$n=(10-1)=9$		$\dot{n}=(10-1)=9$		C = 6

Here, 2C– n, *i.e.*, $2\times6-9=3$ is positive, therefore we use positive (+) sign in the formula. Thus,

$$r_c = \pm\sqrt{\pm\frac{(2C-n)}{n}}$$

$$r_c = \pm\sqrt{\pm\frac{(2\times6-9)}{9}} = +\sqrt{+\frac{3}{9}} = 0.577$$

Thus, there is positive correlation between X and Y.

Example 43. Compute the coefficient of correlation for the following data by the concurrent deviation method:

Year:	1971	1972	1973	1974	1975	1976	1977
Demand:	150	154	160	172	160	165	180
Price:	200	180	170	160	190	180	172

Solution: Denoting Demand and Prices by X and Y.

Year	Demand X	Deviation signs (*dx*)	Price Y	Deviation signs (*dy*)	*dxdy*
1971	150		200		
1972	154	+	180	–	–
1973	160	+	170	–	–
1074	172	+	160	–	–
1075	160	–	190	+	–
1976	165	+	180	–	–
1977	180	+	172	–	–
$n=(7-1)=6$					C = 0

Here, $2C - n$, *i.e.*, $2 \times 0 - 6 = -6$ is negative, therefore we use negative (–) sign in the formula. Thus,

$$r_c = \pm\sqrt{\pm\frac{(2C-n)}{n}}$$

$$= \pm\sqrt{\pm\frac{(2\times 0-6)}{6}} = -\sqrt{-(-1)} = -1$$

There is perfect negative correlation between price and demand.

Example 44. Calculate coefficient of correlation by concurrent deviation method from the following data:

X	112	125	126	118	118	121	125	125	131	135
Y	106	102	102	104	98	96	97	97	95	90

Solution:

Calculation of Coefficient of Concurrent Deviation

X	Deviation signs (dx)	Y	Deviation signs (dy)	$dxdy$
112		106		
125	+	102	–	–
126	+	102	0	0
118	–	104	+	–
118	0	98	–	0
121	+	96	–	–
125	+	97	+	+
125	0	97	0	0
131	+	95	–	–
135	+	90	–	–
$n = 10 - 1 = 9$				$C=2$

Here, $2C - n$, *i.e.*, $2 \times 2 - 9 = -5$ is negative, therefore we use negative (–) sign in the formula. Thus,

$$r_c = \pm\sqrt{\pm\frac{(2C-n)}{n}} \;;\; C = 2,\; n = 9$$

$$= \pm\sqrt{\pm\frac{(2\times 2-9)}{9}} = -\sqrt{-\left(\frac{-5}{9}\right)}$$

$$= -\sqrt{0.5556} = -0.75$$

Thus, there is high degree of negative correlation between X and Y.

IMPORTANT TYPICAL EXAMPLE

Example 45. During the first 9 months of the financial year 1999-2000, the following changes in the price index of shares A and B were recorded as below. Calculate the coefficient of correlation by a suitable method:

Changes over the previous month

Month:	April	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Share A:	–4	–3	–4	0	+3	+4	+2	–3	+3
Share B:	+3	–3	–2	–4	–3	–4	0	–2	–3

Solution: In this question changes are given in comparison to preceding month and in such a case only concurrent deviation method may be used. The value of 'C' will be calculated on the basis of multiplication of signs only (values will be ignored)

Calculation of Coefficient of Correlation

Months	Share A	Deviation signs (dx)	Share B	Deviation signs (dy)	$dxdy$
April	–4	–	+3	+	–
May	–3	–	–3	–	+
June	–4	–	–2	–	+
July	0	0	–4	–	0
August	+3	+	–3	–	–
September	+4	+	–4	–	–
October	+2	+	0	0	0
November	–3	–	–2	–	+
December	+3	+	–3	–	–
	$n = 9$				$C = 3$

Applying the formula,

$$r_c = \pm\sqrt{\pm\frac{(2C-n)}{n}}$$

Here, $C = 3, n = 9$

$$\therefore \quad r_c = \pm\sqrt{\pm\frac{(2\times 3-9)}{9}}$$

$$= \pm\sqrt{\pm\frac{(-3)}{9}} = -\sqrt{-\frac{(-3)}{9}} = -\sqrt{0.33} = -0.574$$

Note: Generally, the value of 'n' is written on the basis of N–1, but in the above example, it will not be applicable because deviation sign of first item is also known.

▶ Merits and Demerits of Concurrent Deviation Method

Merits

(1) This method is simple to understand.
(2) Its computations involve less time.
(3) When the number of items is very large, we can use this method to have a quick idea about the correlation.
(4) This method is useful in studying short term fluctuations.

Demerits

(1) By applying this method, we can get an idea only about the direction of correlation.
(2) This method is not useful for finding correlation of long term changes.
(3) This method is less accurate than Karl Pearson's method.

EXERCISE 1.10

1. Calculate the coefficent of correlation by the method of concurrent deviation from the following data:

X:	65	50	35	55	60	25	45	80	85
Y:	45	35	55	40	70	30	40	65	80

[Ans. $r_c = 0.707$]

2. Calculate coefficient of concurrent deviation from the following data:

X:	65	40	35	75	63	80	35	20	80	60	50
Y:	60	55	50	56	30	70	40	35	80	75	80

[Ans. $r_c = +0.81$]

3. Find coefficient of correlation by concurrent deviation method of the following data:

Students:	A	B	C	D	E	F	G	H
Marks in Economics:	70	45	40	80	68	85	40	25
Marks in Statistics:	65	60	55	61	35	75	45	40

[Ans. $r_c = +1$]

4. Obtain a suitable measure of correlation from the following data regarding changes in price index of two shares A and B during the year:

Changes over the Previous Month

	J	F	M	A	M	J	J	A	S	O	N	D
Shares A:	+4	+3	+2	−1	−3	+4	−5	+1	+2	−7	+2	−3
Shares B:	−2	+5	+3	−2	−1	−3	+4	−1	−3	+6	+4	+2

[Ans. $r_c = -0.48$]

5. Find out the coefficient of correlation between X and Y by the method of concurrent deviations:

X:	26	30	30	24	29	25	25	32	32	38
Y:	62	58	55	68	67	64	64	75	81	78

[Ans. $r_c = -0.577$]

▶ COEFFICIENT OF DETERMINATION

The concept of coefficient of determination is used for the interpretation of coefficient of correlation and comparing the two or more correlation coefficients. **The coefficient of determination is defined as the square of the coefficient of correlation.** It is denoted by r^2. The coefficient of determination explains the percentage variation in the dependent variable Y that can be explained in terms of the independent variable X. If correlation coefficient (r) is 0.9 then coefficient of determination (r^2) will be 0.81 which implies that 81% of the total variations in the dependent variable (Y) occurs due to the independent variable (X). The remaining 19% variation occurs due to outside or external factors. Thus, the coefficient of determination is defined as the ratio of the explained variance to the total variance. In terms of formula:

$$\text{Coefficient of Determination } (r^2) = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

Coefficient of Non-Determination: By dividing the unexplained variation by the total variation, the coefficient of non-determination can be determined. Assuming the total of variation as 1, then the coefficient of determination can be determined by subtracting the coefficient of determination from 1. It is denoted by K^2. In terms of formula,

$$\text{Coefficient of non-determination } (K^2) = 1 - r^2$$

In the above example $r^2 = 0.81$, then the coefficient of non-determination will be 0.19 (1 − 0.81). It indicates that 19% of the variations are due to other factors.

$$\textbf{Coefficient of Alientation} = \sqrt{1 - r^2}$$

Generally, the coefficient of determination $(r)^2$ is widely used in practice.

Example 46. The coefficient of correlation (r) between consumption expenditure (C) and disposable income (Y) in a study was found to be +0.8. What percentage of variation in C are explained by variation in Y?

Solution: Here, $r = 0.8 \Rightarrow r^2 = (0.8)^2 = 0.64$. It means that 0.64 or 64% of the variation in consumption expenditure are explained by variation in income.

Example 47. Is it true that a correlation coefficient (r) = 0.8 indicates a relationship twice as close as $r = 0.4$?

Solution: The statement can be verified by using coefficient of determination, *i.e.*, r^2.

Now, **Ist case:** $r^2 = (0.8)^2 = 0.64$

2nd case: $r^2 = (0.4)^2 = 0.16$

This shows that 64% of the variation is explained in the first case and 16% of the variation is explained in the second case. Hence $r = 0.8$ does not indicate a relationship twice as close as $r = 0.4$.

Example 48. A correlation coefficient of 0.5 implies that 50% of the data are explained. Comment.

Solution: Coefficient of determination (r^2) show the percentage of variation in Y which is explained by the variation in X.

Now, $r^2 = (0.5)^2 = 0.25$

Thus, the coefficient of correlation of 0.5 shows that 25% of the data are explained by X. In other words, 25% of the variation in Y is due to X and the remaining variation is due to other factors.

Example 49. The data relating to import price (X) and import quantity (Y) in respect of a given commodity are as under:

Year:	1975	'76	'77	'78	'79	'80	'81	'82	'83	'84
Import price :	2	3	6	5	4	3	5	7	8	7
Quantity imported:	6	5	4	5	7	10	9	7	8	9

(*i*) Calculate Karl Pearson's coefficient of correlation.

(*ii*) Find the percentage of variation in quantity imported that is explained by the variation in the import price.

Solution: (*i*)

Calculation of Coefficient of Correlation

X	$\bar{X} = 5$ $X - \bar{X}$ x	x^2	Y	$\bar{Y} = 7$ $Y - \bar{Y}$ y	y^2	xy
2	–3	9	6	–1	1	3
3	–2	4	5	–2	4	4
6	+1	1	4	–3	9	–3
5	0	0	5	–2	4	0
4	–1	1	7	0	0	0
3	–2	4	10	+3	9	–6
5	0	0	9	+2	4	0
7	+2	4	7	0	0	0
8	+3	9	8	+1	1	3
7	+2	4	9	+2	4	4
$N = 10$ $\Sigma Y = 50$	$\Sigma x = 0$	$\Sigma x^2 = 36$	$\Sigma Y = 70$	$\Sigma y = 0$	$\Sigma y^2 = 36$	$\Sigma xy = 1$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{50}{10} = 5 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{70}{10} = 7$$

Since the actual means of X and Y are whole numbers, we should take deviations from actual means of X and Y to simplify the calculations.

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{5}{\sqrt{36}\ \sqrt{36}} = \frac{5}{36} = 0.1389$$

(*ii*) Here, $r = 0.1389$

$\Rightarrow$ r^2 = coefficient of determination = $(0.1389)^2 = 0.0192$ or 1.92%

It means that 1.92% of the variations in quantity imported are explained by the variations in the import price.

EXERCISE 1.11

1. The relationship between consumption (C) and disposable income (Y) is expressed by $C = a + by$. In this context, explain what the value of r^2 measures.
2. "A correlation coefficient of 0.3 implies that 30% of the data are explained." Comment.
3. A correlation coefficient of 0.6 indicates a relationship twice as close to as where $r = 0.3$. Comment.

Quantity (Y) :	69	76	52	56	57	77	58	55	67	63	72	64
Price (X) :	9	12	6	10	9	10	7	8	12	6	11	8

(*i*) Calculate the Karl Pearson's coefficient of correlation between price and quantity.

(*ii*) Find the percentage of variation in quantity demanded that is explained by variation in the price of the commodity. [**Ans.** (*i*) $r = 0.645$, (*ii*) 42%]

4. Calculate from the given information:

X:	45	70	65	30	90	40	50	75	75	85	60
Y:	35	90	70	40	95	40	60	80	80	80	50

(*i*) Karl Pearson's coefficient of correlation.

(*ii*) Probable Error and show whether 'r' is significant or not?

(*iii*) Coefficient of non-determination and coefficient of alientation.

[**Ans.** (*i*) $r = 0.904$, (*ii*) P.E. = 0.0390, r is significant, (*iii*) $1 - r^2 = 0.183, 0.4277$]

MISCELLANEOUS SOLVED EXAMPLES

Example 50. (*i*) Find out the coefficient of correlation between X and Y from the following data:

X:	2	2	4	5	5
Y:	6	3	2	6	4

(*ii*) Multiply each X value by 2 and add 3. Multiply each value of Y by 5 and subtract 4. Find the correlation coefficient between two new sets of values. Explain why do or do not obtain the same result as in (*i*).

Solution: (*i*) **Calculation of Karl Pearson's Coefficient of Correlation**

X	X^2	Y	Y^2	XY
2	4	6	36	12
2	4	3	9	6
4	16	2	4	8
5	25	6	36	30
5	25	4	16	20
$\Sigma X = 18$ $N = 5$	$\Sigma X^2 = 74$	$\Sigma Y = 21$	$\Sigma Y^2 = 101$	$\Sigma XY = 76$

$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\ \sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{5\times 76 - 18\times 21}{\sqrt{5\times 74 - (18)^2}\sqrt{5\times 101 - (21)^2}} = 0.036$$

(*ii*) Let us define new variables U and V as follows:

$U = 2X + 3$ and $V = 5Y - 4$

We now calculate the coefficient of correlation between two new sets of values U and V as given

X	Y	$U = 2X + 3$	$V = 5Y - 4$	U^2	V^2	UV
2	6	7	26	49	676	182
2	3	7	11	49	121	77
4	2	11	6	121	36	66
5	6	13	26	169	676	338
5	4	13	16	169	256	208
		$\Sigma U = 51$	$\Sigma V = 85$	$\Sigma U^2 = 557$	$\Sigma V^2 = 1765$	$\Sigma UV = 871$

$$r = \frac{N.\Sigma UV - \Sigma U.\Sigma V}{\sqrt{N.\Sigma U^2 - (\Sigma U)^2}\ \sqrt{N.\Sigma V^2 - (\Sigma V)^2}}$$

$$= \frac{5\times 871 - (51)(85)}{\sqrt{5\times 557 - (51)^2}\sqrt{5\times 1765 - (85)^2}}$$

$$= \frac{20}{\sqrt{184}\sqrt{1600}}$$

$$= 0.036$$

The value of r_{UV} is the same as that of r_{XY}. This is so because the correlation coefficient is independent of the change of origin and scale and U and V are obtained from X and Y by change of origin and scale so that we have r_{XY} and r_{UV}.

Example 51. Two variates X and Y when expressed as deviations from their respective means are given as follows:

x:	0	–4	4	–2	2
y:	1	3	?	0	–1

Find the Karl Pearson Coefficient of correlation between them.

Solution: In this question, one deviation in y series is missing. Let us denote the missing item by a. We know that the sum of deviations taken from mean is always zero.

$$\Sigma y = 0$$

So,

$$\therefore\ (1) + (3) + a + (0) + (-1) = 0$$

$$3 + a = 0$$

$$a = -3$$

$\therefore$

Thus the complete series is:

x:	0	–4	4	–2	2
y:	1	3	–3	0	–1

Now, we find the coefficient of correlation.

Calculation of Coefficient of Correlation

x	x^2	y	y^2	xy
0	0	1	1	0
–4	16	3	9	–12
4	16	–3	9	–12
–2	4	0	0	0
2	4	–1	1	–2
$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

Applying the formula:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{-26}{\sqrt{40\times 20}}$$

$$= \frac{-26}{\sqrt{800}} = \frac{-26}{28.2843} = -0.9192$$

Example 52. The following table gives the distribution of production and also the relatively defective items among them, according to size groups. Find the correlation coefficient between size and defect in quality and its probable error. Is the value of 'r' significant or not?

Size group:	15—16	16—17	17—18	18—19	19—20	20—21
No. of items:	200	270	340	360	400	300
No. of defective items:	150	162	170	180	180	114

Solution: In this question, as correlation has to be found between size and defect in quality, hence defect in quality has to be first determined as % of the defective items.

Calculation of % of Defective Items

No. of items:	200	270	340	360	400	300
No. of defective items:	150	162	170	180	180	114
% of defective items:	$\frac{150}{200}\times 100$ =75	$\frac{162}{270}\times 100$ =60	$\frac{170}{340}\times 100$ =50	$\frac{180}{360}\times 100$ =50	$\frac{180}{400}\times 100$ =45	$\frac{114}{300}\times 100$ =38

Let us denote the mid value of the size group by X and % of defective items as Y

X (MV)	A=18.5 dx	dx^2	Y	A=50 dy	dy^2	$dxdy$
15.5	−3	9	75	25	625	−75
16.5	−2	4	60	10	100	−20
17.5	−1	1	50 = A	0	0	0
18.5 = A	0	0	50	0	0	0
19.5	+1	1	45	−5	25	−5
20.5	+2	4	38	−12	144	−24
$N=6$	$\Sigma dx=-3$	$\Sigma dx^2=19$		$\Sigma dy=18$	$\Sigma dy^2=894$	$\Sigma dxdy$ =−124

Applying the formula,

$$r=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

$$=\frac{6\times(-124)-(-3)(18)}{\sqrt{6\times19-(-3)^2}\sqrt{6\times894-(18)^2}}$$

$$=\frac{-744+54}{\sqrt{114-9}\sqrt{5364-324}}=\frac{-690}{\sqrt{105}\sqrt{5040}}=\frac{-690}{727.46}$$

$$=-0.948\approx-0.95$$

Probable Error (P.E.)

$$PE=0.6745\times\frac{1-r^2}{\sqrt{N}}$$

$$=0.6745\times\left(\frac{1-(-0.95)^2}{\sqrt{6}}\right)$$

$$=0.6745\times\frac{0.0975}{2.45}$$

$$=0.027$$

Significance of 'r'

$$\frac{|r|}{P.E.}=\frac{0.95}{0.027}=35.18$$

$\therefore\ |r|=35.18\ P.E.$

As the value of $|r|$ is more than 6 times the P.E., so 'r' is highly significant.

Example 53. Calculate correlation coefficient from the following results:
$N=10,\ \Sigma X=140,\ \Sigma Y=150$
$\Sigma(X-10)^2=180,\ \Sigma(Y-15)^2=215$
$\Sigma(X-10)(Y-15)=60$

Solution: For calculating correlation coefficient we need the values of $\Sigma X^2, \Sigma Y^2, \Sigma XY$ which we can determine from the values given:

$$\Sigma(X-10)^2=\Sigma(X^2+100-20X)=\Sigma X^2+\Sigma 100-20\Sigma X$$
$$=\Sigma X^2+N\times100-20\Sigma X$$
$$=\Sigma X^2+1000-20\times140 \qquad [\because \Sigma a=Na]$$
$$=\Sigma X^2+1000-2800=\Sigma X^2-1800$$

$\Rightarrow\quad \Sigma X^2-1800=180 \qquad [\because \Sigma(X-10)^2=180]$

$\therefore\quad \Sigma X^2=1980$

$$\Sigma(Y-15)^2=\Sigma(Y^2+225-30Y)=\Sigma Y^2+\Sigma225-30\Sigma Y$$
$$=\Sigma Y^2+N\times225-30\Sigma Y$$
$$=\Sigma Y^2+2250-30\times150 \qquad [\because \Sigma a=Na]$$
$$=\Sigma Y^2+2250-4500=\Sigma Y^2-2250$$

$\Rightarrow\quad \Sigma Y^2-2250=215 \qquad [\because \Sigma(Y-15)^2=215]$

$\therefore\quad \Sigma Y^2=2465$

$$\Sigma(X-10)(Y-15)=\Sigma(XY-15X-10Y+150)$$
$$=\Sigma XY-15\Sigma X-10\Sigma Y+\Sigma150$$
$$=\Sigma XY-15\Sigma X-10\Sigma Y+N\times150$$
$$=\Sigma XY-15\times140-10\times150+10\times150$$
$$=\Sigma XY-2100-1500+1500$$
$$=\Sigma XY-2100$$

$\Rightarrow\quad \Sigma XY-2100=60 \qquad [\because \Sigma(X-10)(Y-15)=60]$

$\therefore\quad \Sigma XY=2160$

Applying the formula,

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

$$=\frac{10\times2160-140\times150}{\sqrt{10\times1980-(140)^2}\sqrt{10\times2465-(150)^2}}$$

$$=\frac{21600-21000}{\sqrt{19800-19600}\sqrt{24650-22500}}$$

$$=\frac{600}{\sqrt{200\times2150}}=\frac{600}{655.74}=+0.915$$

Aliter: r can also be calculated in the following manner:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{140}{10} = 14,\ \bar{Y} = \frac{\Sigma Y}{N} = \frac{150}{10} = 15$$

Thus, deviations $\Sigma(X-10)$ and $\Sigma(Y-15)$ are from assumed means (A_x=10 and A_y=15)

Let, $\Sigma dx = \Sigma(X-10) = \Sigma X - \Sigma 10 = \Sigma X - N.10 = 140 - 10\times 10 = 40$

$\Sigma dy = \Sigma(Y-15) = \Sigma Y - \Sigma 15 = \Sigma Y - N.15 = 150 - 15\times 10 = 0$

$\Sigma\, dxdy = \Sigma(X-10)(Y-15) = 60,\ \Sigma dx^2 = 180,\ \Sigma dy^2 = 215$ (Given)

Applying the formula,

$$r = \frac{N.\Sigma dxdy - \Sigma dx\,.\,\Sigma dy}{\sqrt{N.\Sigma dx^2 - (\Sigma dx)^2}\ \sqrt{N.\Sigma dy^2 - (\Sigma dy)^2}}$$

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

$$= \frac{600}{\sqrt{200\times 2150}} = \frac{600}{655.744} = 0.915$$

Example 54. In two sets of variables X and Y with 50 items each, the following data were observed:
$\bar{X}=10,\ \sigma_x=3, \bar{Y}=6, \sigma_y=2, r=0.3, N=50$

However, on subsequent verification it was found that one value of X(=10) and one value of Y(= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of correlation coefficient affected?

Solution: Given: $N=50,\ \bar{X}=10,\ \bar{Y}=6,\ \sigma_x=3,\ \sigma_y=2,\ r=0.3$

$\therefore\quad \bar{X} = \frac{\Sigma X}{N} \Rightarrow \Sigma X = N\bar{X}$

$\therefore\quad$ Incorrected $\Sigma X = N\,\bar{X} = 50\times 10 = 500$

$\therefore\quad \bar{Y} = \frac{\Sigma Y}{N} \Rightarrow \Sigma Y = N\bar{Y}$ ...(i)

$\therefore\quad$ Incorrected $\Sigma Y = N\,\bar{Y} = 50\times 6 = 300$

$\sigma_x^2 = \frac{\Sigma X^2}{N} - (\bar{X})^2$ ...(ii)

$\Rightarrow\quad \Sigma X^2 = N(\sigma_x^2 + \bar{X}^2)$ [Formula of variance of X]

Incorrected $\Sigma X^2 = N(\sigma_x^2 + \bar{X}^2) = 50(9+100) = 5450$ ...(iii)

$\sigma_y^2 = \frac{\Sigma Y^2}{N} - (\bar{Y})^2$

$\Rightarrow\quad \Sigma Y^2 = N(\sigma_y^2 + \bar{Y}^2)$ [Formula of variance of Y]

Incorrected $\Sigma Y^2 = N(\sigma_y^2 + \bar{Y}^2) = 50(4+36) = 2000$ ...(iv)

We know, $r = \frac{Cov.(X,Y)}{\sigma_x.\sigma_y}$

$\Rightarrow\quad r.\sigma_x.\sigma_y = Cov.(x,y)$

$\because\quad Cov.(X,Y) = \frac{1}{N}.\Sigma(X-\bar{X})(Y-\bar{Y}) = \frac{1}{N}\Sigma XY - \bar{X}\bar{Y}$

$\therefore\quad r.\sigma_x.\sigma_y = \frac{1}{N}.\Sigma XY - \bar{X}.\bar{Y}$

$\therefore\quad \Sigma XY = N[r.\sigma_x.\sigma_y + \bar{X}\bar{Y}]$

$\Sigma XY = 50[0.3\times 3\times 2 + 10\times 6]$

$= 50\,[1.8 + 60]$

$= 50\,[61.8] = 3090$

Incorrected $\Sigma XY = 3090$

Thus, we have the following incorrect values:

$\Sigma X = 500,\ \Sigma Y = 300,\ \Sigma X^2 = 5450,\ \Sigma Y^2 = 2000\ \Sigma XY = 3090$

After dropping out the incorrect values, the corrected values for the remaining 49 pairs of items are given as:

Corrected values:

Corrected $\Sigma X = 500 - 10 = 490$

Corrected $\Sigma Y = 300 - 6 = 294$

Corrected $\Sigma X^2 = 5450 - 10^2 = 5350$

Corrected $\Sigma Y^2 = 2000 - 6^2 = 1964$

Corrected $\Sigma XY = 3090 - 10\times 6 = 3030$

$N = 49$

Using these corrected values, we get

$$r = \frac{\Sigma XY - \frac{\Sigma X.\Sigma Y}{N}}{\sqrt{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}\ .\ \sqrt{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}}$$

$$= \frac{3030 - \frac{490\times 294}{49}}{\sqrt{5350 - \frac{(490)^2}{49}}\ .\ \sqrt{1964 - \frac{(294)^2}{49}}}$$

$$= \frac{3030 - 2940}{\sqrt{450}\ \sqrt{200}} = \frac{90}{300} = +0.3$$

Hence the correlation coefficient is unaffected in this case.

Example 55. "If two variables are independent, the correlation between them is zero, but the converse is not always true." Comment.

Solution: If X and Y are two independent variables, then the covariance between them is zero, i.e., $Cov(X,Y)=0$ and hence $r_{xy}=\frac{Cov(X,Y)}{\sigma_x.\sigma_y}=0$. Thus, if X and Y are independent, they are uncorrelated.

The converse of this property implies that if $r_{xy}=0$, then X and Y may not necessarily be independent. To prove this property, let the two variables X and Y are connected by the relation $Y=X^2$ and consider the following data:

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

Here, $\Sigma X=0, \Sigma Y=28$ and $\Sigma XY=0$

$$\therefore \quad Cov(X,Y)=\frac{1}{N}\Sigma XY-\frac{\Sigma X}{N}\cdot\frac{\Sigma Y}{N}=\frac{1}{7}\cdot(0)-\frac{0}{7}\cdot\frac{28}{7}=0$$

Thus, $$r_{xy}=\frac{Cov(X,Y)}{\sigma_x.\sigma_y}=0$$

A close examination of the data would reveal that although $r_{xy}=0$ but X and Y are not independent. In fact, the variables are related by the equation $Y=X^2$, *i.e.*, there is a quadratic relation (*i.e.*, non-linear relationship) between the variables. This property implies that r_{xy} is only a measure of the linear relationship between X and Y. If the relationship is non-linear, the computed value of r_{xy} is no longer a measure of the degree of relationship between the two variables.

IMPORTANT FORMULAE

A. INDIVIDUAL SERIES

1. Karl Pearson's Coefficient of Correlation (When deviations are taken from actual mean:

$$r=\frac{\Sigma xy}{\sqrt{\Sigma x^2\times\Sigma y^2}} \quad or \quad \frac{\Sigma xy}{N.\sigma_x.\sigma_y}$$

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

2. When deviations are taken from assumed mean:

$$r=\frac{N.\Sigma dxdy-\Sigma dx.\Sigma dy}{\sqrt{N.\Sigma dx^2-(\Sigma dx)^2}\sqrt{N.\Sigma dy^2-(\Sigma dy)^2}}$$

Where, $dx=(X-A)$ and $dy=(Y-A)$

3. When we use actual values of X and Y:

$$r=\frac{N.\Sigma XY-\Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2-(\Sigma X)^2}\sqrt{N.\Sigma Y^2-(\Sigma Y)^2}}$$

4. When we are given Variance and Covariance of X and Y:

$$r=\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

where, $Cov(X,Y)=\frac{1}{N}.\Sigma(X-\bar{X})(Y-\bar{Y})=\frac{1}{N}.\Sigma XY-\bar{X}.\bar{Y}$

B. GROUPED SERIES

5. In a Bivariate or Grouped Frequency Distribution:

$$r=\frac{N\Sigma fdxdy-\Sigma fdx\Sigma fdy}{\sqrt{N.\Sigma fdx^2-(\Sigma fdx)^2}\sqrt{N.(\Sigma fdy)^2-(\Sigma fdy)^2}}$$

6. Spearmen's Rank Correlation Coefficient:

(*i*) When actual ranks are given:

$$R=1-\frac{6\Sigma D^2}{N^3-N}$$

(*ii*) When ranks are not repeated

$$R=1-\frac{6\Sigma D^2}{N^3-N}$$

(*iii*) When ranks are repeated

$$R=1-\frac{6\left[\Sigma D^2+\frac{1}{12}(m^3-m)+\frac{1}{12}(m^3-m)+\ldots\right]}{N^3-N}$$

7. Concurrent Deviation Method

$$r_c=\pm\sqrt{\pm\left(\frac{2C-n}{n}\right)}$$

8. Probable Error and Standard Error

$$P.E.=0.6745\times\frac{1-r^2}{\sqrt{N}} \qquad S.E._r=\frac{1-r^2}{\sqrt{N}}$$

9. Coefficient of Determination

$$r^2=\frac{\text{Explained Variance}}{\text{Total Variance}}$$

QUESTIONS

1. Define correlation. Explain the various methods of studying correlation. What is the significance of studying correlation?
2. What is correlation? Explain various types of correlation. Does it always signify cause and effect relationship between the two variables?
3. Define Pearsons' coefficient of correlation. Interpret r when $r = 1, -1$ and 0.
4. Define rank correlation coefficient. How is it measured? When is it preferred to Karl Pearson's coefficient of correlation?
5. What is meant by coefficient of concurrent deviation? How is it measured?
6. What is scatter diagram and how is it useful in the study of correlation?
7. Explain the followings:
   (*i*) Probable Error (*ii*) Coefficient of Determination.
8. Explain the properties of correlation coefficient.

Linear Regression Analysis

2

INTRODUCTION

The study of regression has special importance in statistical analysis. We know that the mutual relationship between two series is measured with the help of correlation. Under correlation, the direction and magnitude of the relationship between two variables is measured. But it is not possible to make the best estimate of the value of a dependent variable on the basis of the given value of the independent variable by correlation analysis. Therefore, to make the best estimates and future estimation, the study of regression analysis is very important and useful.

MEANING AND DEFINITION

According to **Oxford English Dictionary**, the word 'regression' means "Stepping back" or "Returning to average value". The term was first of all used by a famous Biological Scientist in 19th century, **Sir Francis Galton** relating to a study of hereditary characteristics. He found out an interesting result by making a study of the height of about one thousand fathers and sons. His conclusion was that (*i*) Sons of tall fathers tend to be tall and sons of short fathers tend to be short in height (*ii*) But mean height of the tall fathers is greater than the mean height of the sons, whereas mean height of the short sons is greater than the mean height of the short fathers. The tendency of the entire mankind to **twin** back to average height, was termed by **Galton** 'Regression towards Mediocricity' and the line that shows such type of trend was named as 'Regression Line'.

In statistical analysis, the term 'Regression' is taken in wider sense. **Regression is the study of the nature of relationship between the variables so that one may be able to predict the unknown value of one variable for a known value of another variable.** In regression, one variable is considered as an independent variable and another variable is taken as dependent variable. With the help of regression, possible values of the dependent variable are estimated on the basis of the values of the independent variable. For example, there exists a functional relationship between demand and price, *i.e.*, $D = f(P)$. Here, demand (D) is a dependent variable, and price (P) is an independent variable. On the basis of this relationship between demand and price, probable values of demand can be estimated corresponding to the different values of price.

DEFINITION OF REGRESSION

Some important definitions of regression are as follows:

1. Regression is the measure of the average relationship between two or more variables.
   —**M.M. Blair**
2. Regression analysis measures the nature and extent of the relation between two or more variables, thus enables us to make predictions.
   —**Hirsch**

In brief, regression is a statistical method of studying the nature of relationship between variables and to make prediction.

UTILITY OF REGRESSION

The study of regression is very useful and important in statistical analysis, which is clear by following points:

(1) Nature of Relationship: Regression analysis explains the nature of relationship between two variables.

(2) Estimation of Relationship: The mutual relationship between two or more variables can measured easily by regression analysis.

(3) Prediction: By regression analysis, the value of a dependent variable can be predicted on the basis of the value of an independent variable. For example, if price of a commodity rises, will be the probable fall in demand, this can be predicted by regression.

(4) Useful in Economic and Business Research: Regression analysis is very useful in business and economic research. With the help of regression, business and economic policies can formulated.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION

The main difference between correlation and regression is as follows:

(1) Degree and Nature of Relationship: Correlation is a measure of degree of relationship between X and Y whereas regression studies the nature of relationship between the variables so that one may be able to predict the value of one variable on the basis of another.

(2) Cause and Effect Relationship: Correlation does not always assume cause and effect relationship between two variables. Though two variables may be highly correlated, yet it does not necessary follow that one variable is the cause and another variable is the effect. But regression clearly expresses the cause and effect relationship between two variables. One variable is considered independent in regression, for which the value is given and other variable is dependent which is estimated. The independent variable is the cause and the dependent variable is effect.

(3) Prediction: Correlation does not help in making prediction whereas regression enables us to make prediction. With the help of regression line of Y on X, the probable values of Y can be predicted on the basis of the values of X.

(4) Symmetric: In correlation analysis, correlation coefficient (r_{xy}) is the measure of direction and degree of linear relationship between the two variables X and Y. r_{xy} and r_{yx} are symmetrical i.e., $r_{xy} = r_{yx}$. This implies that it is immaterial which of X and Y is dependent variable and which is independent. In regression analysis, the regression coefficients b_{xy} and b_{yx} are not symmetric, i.e., $b_{xy} \neq b_{yx}$. Thus, correlation coefficients r_{xy} and r_{yx} are symmetric whereas regression coefficients b_{yx} and b_{xy} are not symmetric.

(5) Non-sense Correlation: Sometimes, there may exist spurious or non-sense correlation between two variables by chance, like the correlation, if any between rise in income and rise in weight is a non-sense correlation but in regression analysis, there is nothing like non-sense regression.

(6) Origin and Scale: Correlation coefficient is independent of the change of origin and scale whereas regression coefficient is independent of change of origin but not of scale. This implies that in correlation factor is taken out from X and Y variable, then no adjustment in correlation formula is to be made, whereas in case of regression, we have to make an adjustment for it in our formula.

TYPES OF REGRESSION ANALYSIS

The main types of regression analysis are as follows:

(1) Simple and Multiple Regression: In simple regression analysis, we study only two variables at a time, in which one variable is dependent and another is independent. The functional relationship between income and expenditure is an example of simple regression. On the contrary, we study more than two variables at a time in multiple regression analysis (*i.e.*, at least three variables) in which one is dependent variable and others are independent variable. The study of effect of rain and irrigation on yield of wheat is an example of multiple regression.

(2) Linear and Non-linear Regression: When one variable changes with other variable in some fixed ratio, this is called as linear regression. Such type of relationship is depicted on a graph by means of a straight line or a first degree equation. On the contrary, when one variable varies with other variable in a changing ratio, then it is referred to as curvi-linear/non-linear regression. This relationship, expressed on a graph paper takes the form of a curve. This is presented by way of 2nd or 3rd degree equation.

(3) Partial and Total Regression: When two or more variables are studied for functional relationship but at a time, relationship between only two variables is studied and other variables are hold constant, then it is known as partial regression. On the other hand, in total regression all variables are studied simultaneously for the relationship among them.

SIMPLE LINEAR REGRESSION

In practice, simple linear regression is often used and under this, **Regression Lines, Regression Equations** and **Regression Coefficients** concepts are very important to be studied, which are as follows:

Regression Lines

The regression line shows the average relationship between two variables. This is also known as the **Line of Best Fit.** On the basis of regression line, we can predict the value of a dependent variable on the basis of the given value of the independent variable. If two variables X and Y are given, then there are two regression lines related to them which are as follows:

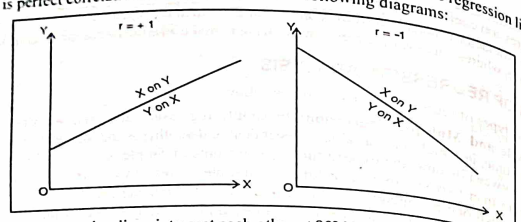
(1) Regression Line of X on Y: The regression line of X on Y gives the best estimate for the value of X for any given value of Y.

(2) Regression Line of Y on X: The regression line of Y on X gives the best estimate for the value of Y for any given value of X.

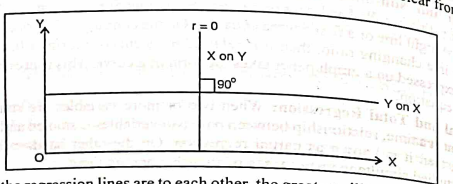
Nature of Regression Lines (or Relation between Correlation and Regression)

With the help of the direction and magnitude of correlation, the nature of regression lines can be known. The main points regarding the relationship among them are as follows:

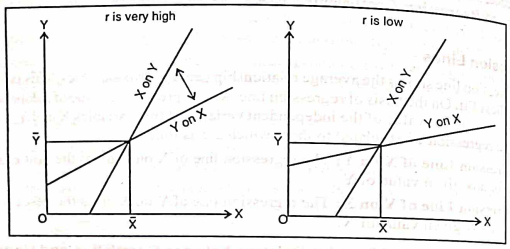
(1) The two regression lines are coincident or there will be only one regression line if $r = \pm 1$, i.e., there is perfect correlation. This is clear from the following diagrams:



(2) The two regression lines intersect each other at 90° if $r = 0$. This is clear from the diagram given below:

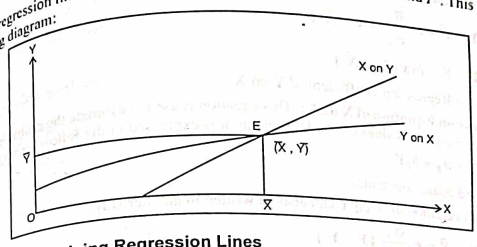


(3) The nearer the regression lines are to each other, the greater will be the degree of correlation. On the contrary, the greater the distance between the two regression lines, the lesser will be the degree of correlation. This is clear from the following diagrams:



(4) If regression lines rise from left to right upward, then correlation is positive. On the other side, if these line move from right to left, then correlation is negative.

(5) The regression lines cut each other at the point of intersection of $\bar{X}$ and $\bar{Y}$. This is clear from the following diagram:



Methods of Obtaining Regression Lines

(1) Scatter Diagram Method,

(2) Least Square Method.

(1) Scatter Diagram Method

This is the simplest method of constructing regression lines. In this method, values of the related variables are plotted on a graph. A straight line is drawn passing through the plotted points. The straight line is drawn with freehand. This shape of regression line can be linear or non-linear also. This depends upon the location of plotted points. This method is very rarely used in practice because in this method, the decision of the person who draws the regression lines very much affects the result.

(2) Least Square Method

Regression lines are also constructed by least square method. Under this method, a regression line is fitted through different points in such a way that the sum of squares of the deviations of the observed values from the fitted line shall be least. The line drawn by this method is called as the **Line of Best Fit.** In other words, under this method, the two regression lines, are drawn in such a way that sum of the squared deviations becomes minimum. The regression line of Y on X is so drawn such that vertically, the sum of squared deviations becomes minimum relating to the different points and the regression line on X on Y is so drawn such that horizontally, squared deviations of different points add up to the minimum.

Regression Equations

Regression equations are the algebraic formulation of regression lines. Regression equations represent regression lines. Just as there are two regression lines, similarly there are two regression equations, which are as follows:

(1) Regression Equation of Y on X: This equation is used to estimate the probable values of Y on the basis of the given values of X. This equation is expressed in the following way:

$$Y = a + bX$$

Here, a and b are constants.

Regression equation of Y on X can also be presented in another way as:

$$Y - \bar{Y} = r \cdot \frac{\sigma_y}{\sigma_x} (X - \bar{X}) .$$

or $Y - \bar{Y} = byx\,(X - \bar{X})$

Here, byx = Regression coefficient of Y on X.

(2) Regression Equation of X on Y: This equation is used to estimate the probable values of X on the basis of the given values of Y. This equation is expressed in the following way:

$$X = a_0 + b_0 Y$$

Here, a_0 and b_0 are constants.

Regression equation of X on Y can also be written in another way:

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

or $X - \bar{X} = bxy\,(Y - \bar{Y})$

Here, bxy = Regression coefficient of X on Y.

○ Regression Coefficients

Just as there are two regression equations, similarly there are two regression coefficients. Regression coefficient measures the average change in the value of one variable for a unit change in the value of another variable. Regression coefficient, in fact, represents the slope of a regression line. For two variables X and Y, there are two regression coefficients, which are given as follows:

(1) Regression Coefficient of Y on X: This coefficient shows that with a unit change in the value of X variable, what will be the average change in the value of Y variable. This is represented by byx. Its formula is as follows:

$$byx = r \cdot \frac{\sigma_y}{\sigma_x}$$

The value of byx can also be determined by other formulae.

(2) Regression Coefficient of X on Y: This coefficient shows that with a unit change in the value of Y variable, what will be the average change in the value of X-variable. It is represented by bxy. Its formula is as follows:

$$bxy = r \cdot \frac{\sigma_x}{\sigma_y}$$

The value of bxy can also be found out by other formulae.

Properties of Regression Coefficients

The main properties of the regression coefficients are as follows:

(1) Coefficient of correlation is the geometric mean of the regression coefficients, *i.e.*,

$$r = \sqrt{bxy \times byx}$$

This property can be proved in the following manner:

Regression coefficient of X on Y $(bxy) = r \cdot \frac{\sigma_x}{\sigma_y}$ ...(i)

Regression coefficient of Y on X $(byx) = r \cdot \frac{\sigma_y}{\sigma_x}$ ...(ii)

Multiplying (i) and (ii)

$$bxy \cdot byx = r \cdot \frac{\sigma_x}{\sigma_y} \cdot r \cdot \frac{\sigma_y}{\sigma_x}$$

or $r^2 = bxy \cdot byx$

Hence, $r = \pm\sqrt{bxy \cdot byx}$

(2) Both the regression coefficients must have the same algebraic signs. The means either both regression coefficients will be either positive or negative. In other words, when one regression coefficient is negative, the other would be also negative. It is never possible that one regression coefficient is negative while the other is positive.

(3) The coefficient of correlation will have the same sign as that of regression coefficients. If both regression coefficient are negative, then the correlation coefficient would be negative. And if byx and bxy have positive signs, then r will also take plus sign.

(4) Both the regression coefficients cannot be greater than unity: If one regression coefficient of y on x is greater than unity, then the regression coefficient of x on y must be less than unity. This is because

$$r = \sqrt{byx \cdot bxy} = \pm 1$$

and never greater than one. If both the regression coefficients happen to be more than 1 then their geometric mean will exceed 1 which will not give the correlation coefficients whose value never exceeds 1.

(5) Arithmetic mean of two regression coefficients is either equal to or greater than the correlation coefficient. In terms of the formula:

$$\frac{byx + bxy}{2} \geq r$$

(6) Shift of origin does not affect regression coefficients but shift in scale does affect regression coefficients. Regression coefficients are independent of the change of origin but not of scale. This means if some common factor is taken out from the items of the series, then in that case, we will have to make adjustment in the regression coefficient formula which is shown below:

$$byx = bvu \cdot \frac{i_y}{i_x} \quad \text{and} \quad bxy = buv \cdot \frac{i_x}{i_y}$$

Where, $u = \frac{X - a}{h}$ and $v = \frac{Y - b}{k}$ and

i_y and i_x are common factors of Y and X series respectively.

To Obtain Regression Equations

The computation of regression equations can be divided into two parts:

(A) Regression Equations in case of Individual Series.

(B) Regression Equations in case of Grouped Data.

▶ (A) Methods to Obtain Regression Equations in case of Individual Series

In individual series, regression equations can be worked out by two methods, which are as follows:

(1) Regression Equations using Normal Equations.

(2) Regression Equations using Regression Coefficients.

▶ (1) Regression Equations using Normal Equations

This method is also called as Least Square Method. Under this method, computation of regression equations is done by solving out two normal equations. This method becomes clear by the following:

Regression Equation of Y on X

Regression Equation of Y on X is expressed as follows:

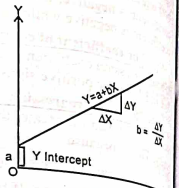
$$Y = a + bX$$

Where, Y = Dependent variable, X = Independent variable,

a = Y-intercept, b = Slope of the line.

Under least square method, the values of a and b are obtained by using the following two normal equations:

$$\Sigma Y = Na + b\Sigma X$$
$$\Sigma XY = a\Sigma X + b\Sigma X^2$$



Solving these equations, we get the following value of a and b.

$$b_{yx} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$$

$$a = \overline{Y} - b\overline{X}$$

Finally, the calculated value of a and b is put in the equation $Y = a + bX$. The regression equation of Y and X will be used to estimate the value of Y when the value of X is given.

Note: a is the Y-intercept, which indicates the minimum value of Y for $X = 0$ and b is the slope of **the line or called regression coefficient of Y and X,** which indicates the absolute increase in Y for a unit increase in X.

Regression Equation of X on Y

Regression Equation of X on Y is expressed as follows:

$$X = a_0 + b_0 Y$$

Where, X = Dependent variable, Y = Independent variable, a_0 = X-intercept, b_0 = Slope of the line

Under least square method, the values of a_0 and b_0 are obtained by using the following two normal equations:

$$\Sigma X = Na_0 + b_0\Sigma Y$$

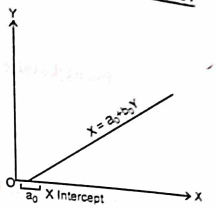
$$\Sigma XY = a_0\Sigma Y + b_0\Sigma Y^2$$

Solving these equations, we get the following value of a_0 and b_0:

$$b_0 = b_{xy} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2}$$

$$a_0 = \overline{X} - b_0\overline{Y}$$

Finally, the calculated value of a_0 and b_0 are put in the equation $X = a_0 + b_0 Y$. The regression equation of X on Y will be used to estimate the value of X when the value of Y is given.



Note: a_0 is the X-intercept, which indicates the minimum value of X for Y = 0 and b_0 **is the slope of the line or called regression coefficient of X on Y.**

The following examples makes the above said method more clear:

Example 1. Calculate the regression equation of X on Y from the following data by the method of least square:

X:	1	2	3	4	5
Y:	2	5	3	8	7

Solution:

Calculation of Regression Equation

X	X^2	Y	Y^2	XY
1	1	2	4	2
2	4	5	25	10
3	9	3	9	9
4	16	8	64	32
5	25	7	49	35
N = 5, $\Sigma X = 15$	$\Sigma X^2 = 55$	$\Sigma Y = 25$	$\Sigma Y^2 = 151$	$\Sigma XY = 88$

Regression Equation of X on Y is

$$X = a + bY$$

The two normal equations are:

$$\Sigma X = Na + b\Sigma Y.$$
$$\Sigma XY = a\Sigma Y + b\Sigma Y^2$$

Substituting the values, we get

$$15 = 5a + 25b \quad \text{...(i)}$$
$$88 = 25a + 151b \quad \text{...(ii)}$$

Multiplying (*i*) by 5 and subtracting it from (*ii*)

$$88 = 25a + 151b$$
$$75 = 25a + 125b$$
$$13 = 26b$$

$\therefore \quad b = \frac{13}{26} = 0.5$

Putting the value of b in equation (i)

$15 = 5a + 25 \times 0.5$

$15 = 5a + 12.5$

$5a = 2.5$

$a = 0.50$

$\therefore \; X = 0.5 + 0.5Y$

Aliter: The value of a and b can also be obtained by using the following formula:

$$b_{xy} = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} \qquad a = \bar{X} - b\bar{Y}$$

Substituting the values, we get

$$b_{xy} = \frac{5 \times 88 - (15)(25)}{5 \times 151 - (25)^2} = \frac{440 - 375}{755 - 625} = \frac{65}{130} = \frac{1}{2} = 0.5$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{15}{5} = 3, \bar{Y} = \frac{\Sigma Y}{N} = \frac{25}{5} = 5$$

$\therefore \quad a = \bar{X} - b\bar{Y} = 3 - \frac{1}{2} \times 5 = 3 - 2.5 = 0.5$

$\therefore \quad X = 0.5 + 0.5Y$

Example 2. Obtain the regression equation of Y on X by the least square method for the following data:

X:	1	2	3	4	5
Y:	9	9	10	12	11

Also estimate the value of Y when X = 10.

Solution:

Calculation of Regression Equation of Y on X

X	Y	XY	X^2
1	9	9	1
2	9	18	4
3	10	30	9
4	12	48	16
5	11	55	25
$N = 5$, $\Sigma X = 15$	$\Sigma Y = 51$	$\Sigma XY = 160$	$\Sigma X^2 = 55$

Regression Equation of Y on X is

$\boxed{Y = a + bX}$

The two normal equations are

$\Sigma Y = Na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

Substituting the values, we get

$51 = 5a + 15b$ ...(i)

$160 = 15a + 55b$ ...(ii)

Multiplying (i) by 3 and subtracting it from (ii)

$160 = 15a + 55b$

$153 = 15a + 45b$

$7 = 10b$

$\therefore \quad b = \frac{7}{10} = 0.7$

Putting the value of b in equation (i)

$51 = 5a + 15(0.7) = 5a + 10.5$

$5a = 40.5$

$a = 8.1$

Hence, the required regression equation of Y on X is given by

$Y = 8.1 + 0.7X$

Estimation for Y

For $X = 10$, $Y = 8.1 + 0.7(10) = 15.1$

Example 3. Given the following data:

$N = 8, \Sigma X = 21, \Sigma X^2 = 99, \Sigma Y = 4, \Sigma Y^2 = 68, \Sigma XY = 36$

Using the values, find

(i) Regression equation of Y on X.

(ii) Regression equation of X on Y.

(iii) Most approximate value of Y for X = 10

(iv) Most approximate value of X for Y = 2.5

Solution: **(i) Regression Equation of Y on X**

$Y = a + bX$

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{8 \times 36 - (21)(4)}{8 \times 99 - (21)^2} = 0.581$$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{21}{8} = 2.625, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{4}{8} = 0.5$$

$\therefore \quad a = \bar{Y} - b\bar{X} = 0.5 - (0.581)(2.625) = -1.025$

$\therefore \quad Y = 1.025 + 0.581X$

(ii) Regression Equation of X on Y

$$X = a_0 + b_0 Y$$

$$b_0 = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{8\times 36 - (21)(4)}{8\times 68 - (4)^2} = 0.386$$

$$a_0 = \bar{X} - b_0\bar{Y} = 2.625 - (0.386)(0.5) = 2.432$$

$\therefore$ $X = 2.432 + 0.386Y$

(iii) Prediction for Y

When $X = 10, Y = -1.025 + 0.581(10) = 4.785$

(iv) Prediction for X

When $Y = 2.5, X = 2.432 + 0.386(2.5) = 3.397$

EXERCISE 2.1

1. Obtain the line of regression of Y on X by least square method for the following data:

X:	1	2	3	4	5
Y:	2	3	5	4	6

Also obtain an estimate of Y when $X = 2$. [Ans. $Y = 1.3 + 0.9X; 3.1$]

2. Find the regression of Y on X and X on Y by the least square method for the following data:

X:	1	2	3
Y:	2	4	5

Also find coefficient of correlation. [Ans. Y= 0.667+1.5X; X= −0.357+0.643Y; r = 0.98]

3. Compute the appropriate regression for the following data:

X (Independent variable):	1	3	4	8	9	11	14
Y (Dependent variable):	1	2	4	5	7	8	9

[Ans. $Y = 0.63X + 0.65$]

4. Obtain the two lines of regression from the following data:
$N = 3, \Sigma X = 6, \Sigma X^2 = 14, \Sigma Y = 15, \Sigma Y^2 = 77, \Sigma XY = 31$

[Ans. $Y = 0.5X + 4$, $X = 0.5Y - 0.5$]

5. Given: $\Sigma Y = 15, \Sigma Y = 110, \Sigma XY = 400, \Sigma X^2 = 250, \Sigma Y^2 = 3200, N = 10$
Find the following:
(*i*) Regression coefficient of Y on X and the Y-intercept.
(*ii*) X-intercept, and the regression coefficient of X on Y.
(*iii*) Most approximate value of Y for X = 5.
(*iv*) Most approximate value of X for Y = 25.

[Ans. (*i*) $b = 1.033$, $a = 9.451$, (*ii*) $a = 0.201$, $b = 0.118$, (*iii*) $Y = 14.616, X = 315$]

(2) Regression Equations using Regression Coefficients

Regression equations can also be computed with the help of regression coefficients. For this, we have to find out $\bar{X}, \bar{Y}, byx$ and bxy from the given data. Regression equations can be constructed from the regression coefficients by any of the following methods:

(*i*) Using the actual values of X and Y series.
(*ii*) Using deviations from Actual Means.
(*iii*) Using deviations from Assumed Means.
(*iv*) Using r, σ_x, σ_y and $\bar{X}, \bar{Y}$.

(i) Using the Actual Values of X and Y Series

In this method, actual values of X and Y are used to determine regression equations. With regard to regression coefficients, regression equations are put in the following way:

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

or $$Y = \bar{Y} + byx\,(X - \bar{X})$$

Here, $\bar{X}$ = Arithmetic mean of X series $= \frac{\Sigma X}{N}$

$\bar{Y}$ = Arithmetic mean of Y series $= \frac{\Sigma Y}{N}$

byx = Regression coefficient of Y on X

Using actual values, the value of byx can be calculated as:

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} \quad \text{or} \quad byx = \frac{\Sigma XY / N - \bar{X}:\bar{Y}}{\sigma_x^2}$$

Note: This formula is based on the normal equations, yet its use avoids the solution of normal equations.

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

or $$X = \bar{X} + bxy(Y - \bar{Y})$$

Where bxy= Regression coefficient of X on Y.

Using actual values, the value of bxy can be calculated as:

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} \quad \text{or} \quad bxy = \frac{\Sigma XY / N - \bar{X}.\bar{Y}}{\sigma_y^2} = \frac{Cov(X,Y)}{\sigma_y^2}$$

The following examples make this method more clear:

Example 4. Calculate the regression equations of X on Y and Y on X from the following data:

X:	1	2	3	4	5
Y:	2	5	3	8	7

Solution:

Calculation of Regression Equations

X	X^2	Y	Y^2	XY
1	1	2	4	2
2	4	5	25	10
3	9	3	9	9
4	16	8	64	32
5	25	7	49	35
N=5, ΣX=15	ΣX^2=55	ΣY=25	ΣY^2=151	ΣXY=88

$$\bar{X} = \frac{\Sigma X}{N} = \frac{15}{5} = 3, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{25}{5} = 5$$

Regression Coefficient of Y on X (by.x):

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{5\times 88 - (15)(25)}{5\times 55 - (15)^2} = \frac{440-375}{275-225} = \frac{65}{50} = 1.3$$

Regression Coefficient of X on Y (bxy):

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2}$$

$$= \frac{5\times 88 - (15)(25)}{5\times 151 - (25)^2} = \frac{440-375}{755-625} = \frac{65}{130} = +0.5$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

Substituting the values,

$$Y - 5 = 1.3(X - 3)$$
$$Y - 5 = 1.3X - 3.9$$
$$Y = 1.3X - 3.9 + 5$$
$$Y = 1.3X + 1.1$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 3 = +0.5\,(Y - 5)$$
$$X - 3 = 0.5Y - 2.5$$
$$X = 0.5Y + 0.5$$

Example 5. Calculate the two regression equations from the following data:

$\Sigma X = 30, \Sigma Y = 23, \Sigma XY = 168, \Sigma X^2 = 224, \Sigma Y^2 = 175, N = 7$

Hence or otherwise find Karl Pearson's coefficient of correlation.

Solution:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{7} = 4.286$$

$$\bar{Y} = = \frac{\Sigma Y}{N} = \frac{23}{7} = 3.286$$

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{7\times 168 - (30)(23)}{7\times 224 - (30)^2} = 0.728$$

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{7\times 168 - (30)(23)}{7\times 175 - (23)^2} = 0.698$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 3.286 = 0.728\,(X - 4.286)$$
$$Y - 3.286 = 0.728X - 3.120$$
$$Y = 0.728X + 0.166$$

∴ **Regression Equation of X on Y**

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 4.286 = 0.698\,(Y - 3.286)$$
$$X - 4.286 = 0.698Y - 2.294$$
$$X = 0.698Y + 1.992$$

∴ **Karl Pearson's Coefficient of Correlation**

$$r = \sqrt{byx \cdot bxy}$$
$$r = \sqrt{0.728\times 0.698} = 0.712$$

IMPORTANT TYPICAL EXAMPLES

Example 6. In order to find the correlation coefficient between the two variables X and Y from 12 pairs of observations, the following calculations were made:

$\Sigma X = 30, \Sigma X^2 = 670, \Sigma Y = 5, \Sigma Y^2 = 285, \Sigma XY = 344$

On subsequent verifications, it was discovered that the pair (X = 11, Y = 4) was copied wrongly, the correct values being (X = 10, Y = 14). After making necessary corrections, find:

(*i*) the two regression coefficients.

(*ii*) the two regresssion equations.

(*iii*) the correlation coefficient.

Solution: Corrected ΣX = 30 + Correct value – Incorrect value

= 30 + 10 – 11 = 29

Corrected ΣY = 5 + 14 – 4 = 15

Corrected $\Sigma Y^2 = 670 + (\text{Correct value})^2 - (\text{Incorrect value})^2$
$= 670 + 10^2 - 11^2 = 649$

Corrected $\Sigma Y^2 = 285 + 14^2 - 4^2 = 465$

Corrected $\Sigma XY = 344 + (10)(14) - (11)(4) = 440$

$$\bar{X} = \frac{\text{Corrected } \Sigma X}{N} = \frac{29}{12} = 2.416$$

$$\bar{Y} = \frac{\text{Corrected } \Sigma Y}{N} = \frac{15}{12} = 1.25$$

(i) Regression Coefficients

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{12\times 440 - 29\times 15}{12\times 649 - (29)^2}$$
$$= \frac{5280-435}{7788-841} = \frac{4845}{6947} = +0.697$$

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{12\times 440 - 29\times 15}{12\times 465 - (15)^2}$$
$$= \frac{5280-435}{5580-225} = \frac{4845}{5355} = 0.904$$

(ii) Two Regression Equations

X on Y

$X - \bar{X} = bxy\,(Y - \bar{Y})$
$X - 2.416 = 0.904\,(Y - 1.25)$
$X - 2.416 = 0.904Y - 1.13$
$X = 0.904Y + 1.286$

Y on X

$Y - \bar{Y} = byx\,(X - \bar{X})$
$Y - 1.25 = 0.697\,(X - 2.416)$
$Y - 1.25 = 0.697X - 1.683$
$Y = 0.697X - 0.433$

(iii) Correlation coefficient

$$r = \sqrt{byx\,.\,bxy} = \sqrt{(0.697)(0.904)} = +0.793$$

Example 7. Given that

$\bar{X} = 15,\ \bar{Y} = 12, \Sigma XY = 1500, \sigma_x = 6.4, \sigma_y = 9.0, N = 10$

Compute: (a) Two regression Coefficients

(b) Correlation coefficient between X and Y.

Solution: Given: $\bar{X} = 15,\ \bar{Y} = 12, \Sigma XY = 1500, \sigma_x = 6.4, \sigma_y = 9.0, N = 10$

Regression Coefficient of Y on X

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2}$$

The values of N and ΣXY are given and the values of $\Sigma X^2, \Sigma Y^2, \Sigma X$ and ΣY are to be calculated as follows:

$$\bar{X} = \frac{\Sigma X}{N} \Rightarrow \Sigma X = N.\bar{X} = 10\times 15 = 150$$

$$\bar{Y} = \frac{\Sigma Y}{N} \Rightarrow \Sigma Y = N.\bar{Y} = 10\times 12 = 120$$

$$\Sigma X^2 = N[\sigma_x^2 + (\bar{X})^2] = 10[6.4^2 + 15^2] = 2659.6$$

$$\Sigma Y^2 = N[\sigma_y^2 + (\bar{Y})^2] = 10[9^2 + 12^2] = 2250$$

Now,
$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} = \frac{10\times 1500 - (150)(120)}{10\times 2659.6 - (150)^2}$$
$$= \frac{15000-18000}{26596-22500} = \frac{-3000}{4096} = -0.73$$

Aliter: byx can also be calculated as follows:

$$byx = \frac{\frac{\Sigma XY}{N} - \bar{X}.\bar{Y}}{\sigma_x^2} = \frac{\frac{1500}{10} - (15)(12)}{(6.4)^2}$$
$$= \frac{150-180}{40.96} = \frac{-30}{40.96} = -0.73$$

Regression Coefficient of X on Y

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} = \frac{10\times 1500 - (150)(120)}{10\times 2250 - (120)^2}$$
$$= \frac{15000-18000}{22500-14400} = \frac{-3000}{8100} = -0.37$$

Aliter: bxy can also be calculated as follows:

$$bxy = \frac{\frac{\Sigma XY}{N} - \bar{X}.\bar{Y}}{\sigma_y^2} = \frac{\frac{1500}{10} - (15)(12)}{(9)^2}$$
$$= \frac{150-180}{81} = \frac{-30}{81} = -0.37$$

Coefficient of Correlation

$$r = \pm\sqrt{byx.bxy} = -\sqrt{(-0.73)\times(-0.37)} = -0.519$$

Example 8. Find out the regression coefficients of Y on X and X on Y from the following data: $\Sigma X = 50, \bar{X} = 5, \Sigma Y = 60, \bar{Y} = 6, \Sigma XY = 350$, Variance of X = 4, Variance of Y = 9.

Solution: We know that: $\bar{X} = \frac{\Sigma X}{N} \Rightarrow 5 = \frac{50}{N} \Rightarrow N = 10$

$$byx = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma X^2 - (\Sigma X)^2} \text{ or } \frac{\Sigma XY/N - \bar{X}.\bar{Y}}{\sigma_x^2} \text{ or } \frac{Cov(X,Y)}{\sigma_x^2}$$

$$\therefore \quad byx = \frac{\frac{350}{10} - (5)(6)}{4} = \frac{35-30}{4} = \frac{5}{4}$$

$$= 1.25$$

$$bxy = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{N.\Sigma Y^2 - (\Sigma Y)^2} \text{ or } \frac{\Sigma XY/N - \bar{X}.\bar{Y}}{\sigma_y^2} \text{ or } = \frac{Cov(X,Y)}{\sigma_y^2}$$

$$\therefore \quad bxy = \frac{\frac{350}{10} - (5)(6)}{9} = \frac{35-30}{9} = \frac{5}{9}$$

EXERCISE 2.2

1. Given the following bivariate data:

X:	−1	5	3	2	1	1	7	3
Y:	−6	1	0	0	1	2	1	5

(*i*) Fit a regression line of Y on X and predict Y if X=10.

(*ii*) Fit a regression line of X on Y and predict X if Y=2.5

[**Ans.** $Y = -1.025 + 0.581X$; $X = 2.432 + 0.386Y$; $Y_{10} = 4.785$; $X_{2.5} = 3.397$]

2. By using the following data, find the regression equation of Y on X and compute the value of Y when X = 10.

$\bar{X} = 5.5$, $\bar{Y} = 4.0$, $\Sigma X^2 = 385$, $\Sigma Y^2 = 192$, $\Sigma(X+Y)^2 = 947$ and $N = 10$

[**Ans.** $Y = -0.42X + 6.31$, $Y_{10} = 2.11$]

3. Given that:

$\Sigma X = 250, \Sigma Y = 300$, $\sigma_x = 5$, $\sigma_y = 10$, $\Sigma XY = 7900$, $N = 10$

Compute : (*i*) Two regression coefficients,

(*ii*) Correlation coefficient between X and Y.

(*iii*) Most approximate value of Y when X = 55 and X when Y = 40.

[**Ans.** $byx = 1.6$, $bxy = 0.4$, $r = 0.8$, $Y_{30} = 78$, $X_{40} = 28$]

4. By using the following data, find correlation coefficient and regression equation of Y on X and estimated value of Y when X = 20

$N = 10, \Sigma X = 140, \Sigma Y = 150, \Sigma(X-10)^2 = 180, \Sigma(Y-15)^2 = 215, \Sigma(X-10)(Y-15) = 60$

[Hint: See Example 53 on Correlation]

[**Ans.** $r = 0.915$, $Y = 3X - 27$, $Y_{20} = 33$]

5. Following information was computed through a computer:

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

Later on it was discovered that two pairs of X and Y were miscopied as (6, 14) and (8, 6) instead of (8, 12) and (6, 8). Determine (*i*) the correct regression equations (*ii*) correct coefficient of correlation. [**Ans.** (*i*) $X = 0.556Y + 2.776$, $Y = 0.8X$, (*ii*) $r = 0.67$]

6. On each of 30 sets, two measurements are made. The following summaries are given :

$\Sigma X = 15, \Sigma Y = -6, \Sigma XY = 56, \Sigma X^2 = 61$ and $\Sigma Y^2 = 90$

Calculate the product moment correlation coefficient and the slope of regression line of Y on X.

[Hint: See Example 52] [**Ans.** $r = 0.856$, $byx = 1.10$]

(*i*) Using Deviations taken from Actual Means

When the size of the values of X and Y is very large, then the method using actual values becomes very difficult to use. In such case, in place of actual values, deviations taken from arithmetic means ($\bar{X}$, $\bar{Y}$) are used to simplify the computation process. In such a case, regression equations are expressed as follows:

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

or $$Y = \bar{Y} + byx\,(X - \bar{X})$$

Here, $\bar{X}$ = Arithmetic mean of X

$\bar{Y}$ = Arithmetic mean of Y

byx = Regression coefficient of Y on X

Using deviations from actual means, the value of byx can be calculated as:

$$\boxed{byx = \frac{\Sigma xy}{\Sigma x^2}}$$

Where, $x = X - \bar{X}$; $y = Y - \bar{Y}$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

or $$X = \bar{X} + bxy(Y - \bar{Y})$$

Where, bxy = Regression coefficient of X on Y.

Using deviations from actual means, the value of bxy can be calculated as:

$$\boxed{bxy = \frac{\Sigma xy}{\Sigma y^2}}$$

Where, $x = X - \bar{X}$; $y = Y - \bar{Y}$

The following examples make this method more clear.

Example 9. Obtain the two regression equations from the following data:

X:	2	4	6	8	10	12
Y:	4	2	5	10	3	6

Solution:

X	$\bar{X}=7$ $(X-\bar{X})$ x	x^2	Y	$\bar{Y}=5$ $(Y-\bar{Y})$ y	y^2	xy
2	−5	25	4	−1	1	+5
4	−3	9	2	−3	9	+9
6	−1	1	5	0	0	0
8	+1	1	10	+5	25	+5
10	+3	9	3	−2	4	−6
12	+5	25	6	+1	1	+5
$\Sigma X=42$ $N=6$	$\Sigma x=0$	$\Sigma x^2=70$	$\Sigma Y=30$	$\Sigma y=0$	$\Sigma y^2=40$	$\Sigma xy=18$

$$\bar{X}=\frac{\Sigma X}{N}=\frac{42}{6}=7;\quad \bar{Y}=\frac{\Sigma Y}{N}=\frac{30}{6}=5$$

Since, the actual means of X and Y are whole numbers, we should taken deviations from $\bar{X}$ and $\bar{Y}$ to simplify calculations:

$$byx=\frac{\Sigma xy}{\Sigma x^2}=\frac{18}{70}=0.257$$

$$bxy=\frac{\Sigma xy}{\Sigma y^2}=\frac{18}{40}=0.45$$

Regression Equation of Y on X

$$Y-\bar{Y}=byx\,(X-\bar{X})$$
$$Y-5=0.257\,(X-7)$$
$$Y-5=0.257X-1.799$$
$$\therefore\quad Y=0.257X+3.201$$

Regression Equation of X on Y

$$X-\bar{X}=bxy\,(Y-\bar{Y})$$
$$X-7=0.45(Y-5)$$
$$X-7=0.45Y-2.25$$
$$X=0.45Y-2.25+7$$
$$\therefore\quad X=0.45Y+4.75$$

Example 10. The following are the intermediate results of the two series X and Y:
$\bar{X}=90, \bar{Y}=70, N=10, \Sigma x^2=6360, \Sigma y^2=2860, \Sigma xy=3900$
(Where x and y are deviations from the respective means)
Find two regression equations.

Solution:

Regression Coefficient of Y on X

$$byx=\frac{\Sigma xy}{\Sigma x^2}=\frac{3900}{6360}=0.613$$

Regression Coefficient of X on Y

$$bxy=\frac{\Sigma xy}{\Sigma y^2}=\frac{3900}{2860}=1.363$$

Regression Equation of X on Y

$$X-\bar{X}=bxy\,(Y-\bar{Y})$$
$$X-90=1.363\,(Y-70)$$
$$X-90=1.363Y-95.41$$
$$X=1.363Y-5.41$$
$$\therefore$$

Regression Equation of Y on X

$$Y-\bar{Y}=byx\,(X-\bar{X})$$
$$Y-70=0.613\,(X-90)$$
$$Y-70=0.613X-55.17$$
$$Y=0.613X+14.83$$
$$\therefore$$

Example 11. The following table gives the aptitude test scores and productivity indices of 10 workers at random:

Aptitude score	Productivity index
60	68
62	60
65	62
70	80
72	85
48	40
53	52
73	62
65	60
82	81

Estimate:
(*i*) the test score of a worker whose productivity index is 75.
(*ii*) the productivity index of a worker whose test score is 92.

Solution:

Calculation of Regression Equations

Aptitude Score X	$(\bar{X}=65)$ x	x^2	Productivity index Y	$(\bar{Y}=65)$ y	y^2	xy
60	−5	25	68	+ 3	9	−15
62	−3	9	60	−5	25	+ 15
65	0	0	62	−3	9	0
70	+ 5	25	80	+ 15	225	+ 75
72	+ 7	49	85	+ 20	400	+ 140
48	−17	289	40	−25	625	+ 425
53	−12	144	52	−13	169	+ 156
73	+ 8	64	62	−3	9	−24
65	0	0	60	−5	25	0
82	+ 17	289	81	+ 16	256	+ 272
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 894$	$\Sigma Y = 650$	$\Sigma y = 0$	$\Sigma y^2 = 1752$	$\Sigma xy = 1044$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{650}{10} = 65 \; ; \; \bar{Y} = \frac{\Sigma Y}{N} = \frac{650}{10} = 65$$

Regression Equation of X on Y: $X - \bar{X} = bxy(Y - \bar{Y})$

$$bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{1044}{1752} = +0.596$$

$$X - 65 = 0.596\,(Y - 65)$$

$$X - 65 = 0.596\,Y - 38.74$$

or $$X = 26.26 + 0.596\,Y$$

For finding out the test score (X) of a person whose productivity index (Y) is 75, put Y = 75 in the above equation:

$$X_{75} = 26.26 + 0.596(75) = 26.26 + 44.7 = 70.96.$$

Regression Equation of Y on X: $Y - \bar{Y} = byx\,(X - \bar{X})$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{1044}{894} = +1.168$$

$$Y - 65 = 1.168\,(X - 65)$$

$$Y - 65 = 1.168\,X - 75.92 \text{ or } Y = -10.92 + 1.168\,X$$

For finding out the productivity index (Y) of a worker whose test score (X) is 92, we put X = 92 in the above equation.

$$Y_{92} = -10.92 + 1.168(92)$$

$$= -10.92 + 107.456 = 96.536$$

IMPORTANT TYPICAL EXAMPLES

Example 12. The following table shows the number of motor registrations in a certain territory for a term of 5 years and the sale of motor tyres by a firm in that territory for the same period:

Year:	1	2	3	4	5
Motor registration:	600	630	720	750	800
No. of tyres sold:	1,250	1,100	1,300	1,350	1,500

Find the regression equation to estimate the sale of tyres when motor registration is known. Estimate the sale of tyres when registration is 850.

Let X denotes number of motor registrations and Y denotes the number of tyres sold by a firm.

Solution: To simplify the calculation, let

$$x = \frac{X - \bar{X}}{i_x} \qquad y = \frac{Y - \bar{Y}}{i_y}$$

X	$x = \frac{X-\bar{X}}{10}$	x^2	Y	$y = \frac{Y-\bar{Y}}{50}$	y^2	xy
600	− 10	100	1,250	− 1	1	+ 10
630	− 7	49	1,100	− 4	16	+ 28
720	2	4	1,300	0	0	0
750	5	25	1,350	+ 1	1	+ 5
800	10	100	1,500	+ 4	16	+ 40
$\Sigma X = 3500$ $N = 5$	$\Sigma x = 0$	$\Sigma x^2 = 278$	$\Sigma Y = 6500$ $N = 5$	$\Sigma y = 0$	$\Sigma y^2 = 34$	$\Sigma xy = 83$

$$\bar{X} = \frac{3500}{5} = 700, \qquad \bar{Y} = \frac{6500}{5} = 1300$$

Here, we have the regression of Y on X.

$$byx = \frac{\Sigma xy}{\Sigma x^2} \times \frac{i_y}{i_x} = \frac{83}{278} \times \frac{50}{10} = 1.4928$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 1300 = 1.4928\,(X - 700)$$

$$Y - 1300 = 1.4928X - 1044.96$$

$$Y = 1.4928X + 255.04$$

The estimate of sale of tyres (Y) when registration $X = 850$ is given by

$$Y = 1.4928 \times 850 + 255.04$$

$$= 1268.88 + 255.04 = 1523.92 \simeq 1524$$

since the number of tyres cannot be fractional.

Example 13. Calculate the correlation coefficient from the following results:

$N = 10, \Sigma X = 350, \Sigma Y = 310$

$\Sigma (X - 35)^2 = 162, \Sigma (Y - 31)^2 = 222, \Sigma (X - 35)(Y - 31) = 92$

Also find the regression line of Y on X.

Solution:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{350}{10} = 35$$

$$\bar{Y} = \frac{\Sigma Y}{N} = \frac{310}{10} = 31$$

Thus, the given deviations $(X - 35)$ and $(Y - 31)$ are from actual means $(\bar{X} = 35,\ \bar{Y} = 31)$.

Thus, $\Sigma (X - 35)^2 = 162$ or $\Sigma x^2 = 162$ where, $x = X - \bar{X}$

$\Sigma (Y - 31)^2 = 222$, or $\Sigma y^2 = 222$ $y = Y - \bar{Y}$

$\Sigma (X - 35)(Y - 31) = 92$ *or* $\Sigma xy = 92$

Coefficient of Correlation

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\ \sqrt{\Sigma y^2}} = \frac{92}{\sqrt{162}\ \sqrt{222}} = +0.485$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{92}{162} = 0.568$$

$\therefore$ $Y - 31 = 0.568(X - 35)$

$Y - 31 = 0.568X - 19.88$

$Y = 0.568X - 19.88 + 31$

$Y = 0.568X + 11.12$

○ Graphing Regression Lines

It is quite easy to graph the regression lines once they have been computed. The procedure adopted is as follows:

***(i)* Regression line of X on Y.** The regression line of X on Y can be drawn with the help of regression equation of X on Y, *i.e.*,

$$X = a + bY$$

If we put the respective values of Y in the above regression equation, we will find the estimated values of X. If we plot estimated values of X with the actual values of Y on the graph, we can draw regression line of X on Y.

***(ii)* Regression line of Y on X.** The regression line of Y on X can be drawn with the help of regression equation of Y on X, *i.e.*,

$$Y = a + bX$$

If we put the respective values of X in the above equation, we will find the estimated values of Y. If we plot estimated values of Y with the actual values of X on the graph, we can draw regression line of Y on X.

The following example illustrate the graphing of regression lines.

Example 14. From the following data:

(i) Obtain the two regression equations.

(ii) Draw up the two regression lines on the graph paper with the help of two regression equations.

X:	1	2	3
Y:	5	4	6

Solution:

Calculation of Regression Equation

X	$\bar{X} = 2$ x	x^2	Y	$\bar{Y} = 5$ y	y^2	xy
1	−1	1	5	0	0	0
2	0	0	4	−1	1	0
3	+1	1	6	+1	1	+1
$\Sigma X = 6$ $N = 3$	$\Sigma x = 0$	$\Sigma x^2 = 2$	$\Sigma Y = 15$ $N = 3$	$\Sigma y = 0$	$\Sigma y^2 = 2$	$\Sigma xy = +1$

$\therefore$ $\bar{X} = \frac{\Sigma X}{N} = \frac{6}{3} = 2$; $\bar{Y} = \frac{15}{3} = 5$

$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{1}{2}$ $bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{1}{2}$

Regression Equation of Y on X

$Y - \bar{Y} = byx\,(X - \bar{X})$

$Y - 5 = \frac{1}{2}(X - 2)$

$Y - 5 = \frac{1}{2}X - 1$

$\therefore$ $Y = \frac{1}{2}X + 4$

Regression Equation of X on Y

$X - \bar{X} = bxy\,(Y - \bar{Y})$

$X - 2 = \frac{1}{2}(Y - 5)$

$X - 2 = \frac{1}{2}Y - \frac{5}{2}$

$\therefore$ $X = \frac{1}{2}Y - \frac{1}{2}$

***(ii)* Regression Lines:** In order to draw up the two regression lines on the graph, we shall have to plot the given values of X and the computed values of Y and the given values of Y and the computed values of X

Computed Values of Y

Regression equation of Y on X

$$Y = \frac{1}{2}X + 4$$

Computed Values of X

Regression equation of X on Y

$$X = \frac{1}{2}Y - \frac{1}{2}$$

when $X = 1$, $Y = \frac{1}{2}(1) + 4 = 4.5$

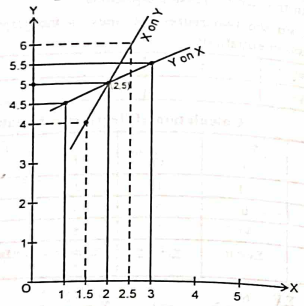
when $X = 2$, $Y = \frac{1}{2}(2) + 4 = 5.0$

when $X = 3$, $Y = \frac{1}{2}(3) + 4 = 5.5$

when $Y = 5$, $X = \frac{1}{2}(5) - \frac{1}{2}$

when $Y = 4$, $X = \frac{1}{2}(4) - \frac{1}{2}$

when $Y = 6$, $X = \frac{1}{2}(6) - \frac{1}{2}$



Example 15. Compute the appropriate regression equation for the following data:

X (Independent variable)	Y (Dependent variable)
2	18
4	12
5	10
6	8
8	7
11	5

Solution: The appropriate regression equation will be Y on X

X	$\bar{X} = 6$ x	x^2	Y	$\bar{Y} = 10$ y	y^2	xy
2	−4	16	18	8	64	−32
4	−2	4	12	2	4	−4
5	−1	1	10	0	0	0
6	0	0	8	−2	4	0
8	+2	4	7	−3	9	−6
11	+5	25	5	−5	25	−25
$\Sigma X = 36$	$\Sigma x = 0$	$\Sigma x^2 = 50$	$\Sigma Y = 60$	$\Sigma y = 0$	$\Sigma y^2 = 106$	$\Sigma xy = -67$

$\bar{X} = \frac{36}{6} = 6$; $\bar{Y} = \frac{60}{6} = 10$

$$byx = \frac{\Sigma xy}{\Sigma x^2}$$

$$= \frac{-67}{50} = -1.34$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 10 = -1.34\,(X - 6)$$

$$Y - 10 = -1.34X + 8.04$$

$$\therefore \quad Y = -1.34X + 18.04$$

EXERCISE 2.3

1. For the following data, set up regression equation and estimate sales for an advertisement expenditure of Rs. 75 lakh.

Sales (Rs. crore):	14	16	18	20	24	30	32
Adv. expenditure (Rs. lakh):	52	62	65	70	76	80	78

[Hint: Let X denote sales) [Ans. X = 0.621Y − 20.85, X_{75} = 25.725]

2. Find the correlation coefficient and the equations of regression lines for the following values of X and Y:

X:	11	7	2	5	8	6	10
Y:	7	5	3	2	6	4	8

[Ans. r = 0.884, X = 0.75 + 1.25Y, Y = 0.625 + 0.625X]

3. The following data relate to marketing expenditure and the corresponding sales:

Expenditure (X) (Rs. lac):	10	12	15	20	23
Sales (Y) (Rs. crore) :	14	17	23	21	35

Estimate the marketing expenditure to obtain a sales target of Rs. 40 crore.
[Ans. X = 0.59Y + 3.02; X_{40} = 26.62]

4. The following are the intermediate results of the two series X and Y
$\bar{X} = 65, \bar{Y} = 65, N = 10, \Sigma x^2 = 894, \Sigma y^2 = 1752, \Sigma xy = 1044$
(Where x and y are deviations from the respective means)
Find two regression equations. Also estimate Y when X = 92 and X when Y = 75.
[Ans. Y = 1.168X − 10.92, Y_{92} = 96.536; X = 0.596Y + 26.26, X_{75} = 70.96]

5. An investigation in to the demand for Television sets in 7 towns has resulted in the follow... data:

Population ('000) (X):	11	14	14	17	17	21	23
No. of T.V. sets demanded (Y):	15	27	27	30	34	38	45

Calculate the regression equation of Y on X and estimate the demand for T.V. sets for a town with a population of 30 thousand. [Ans. $Y = -3 + 2X$; $Y_{30} = 57$]

6. A departmental store gives in-service training to its salesmen which is followed by a test. It is considering whether it should terminate the service of any salesman who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period:

Test scores:	14	19	24	21	26	22	15	20	19
Sales ('00 Rs.):	31	36	48	37	50	45	33	41	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low test scores is justified ? If the firm wants a minimum sales volume of Rs. 3,000, what is the minimum test score that will ensure continuation of service ? Also estimate the most probable sales volume of a salesman making a score of 28.
[Hint : See Example 57] [Ans. $r = 0.9471$, justified, $X = 14.422 \simeq 14$, $Y = 5286.61$]

7. The following table gives the marks in Economics and Statistics of 10 students selected at random:

Marks in Economics:	25	28	35	32	31	36	29	38	34	32
Marks in Statistics:	43	46	49	41	36	32	31	30	33	39

Find (*i*) The two regression equations.
(*ii*) The coefficient of correlation between marks in Economics and Statistics.
(*iii*) The most likely marks in statistics when marks in economics are 30.
[Ans. (*i*) $X = -0.2337Y + 40.8806$, $Y = -0.6643X + 59.257$
(*ii*) $r = -0.394$, (*iii*) 39.3286, or 39 marks]

8. The profits (Y) of a company in the Xth year of its life were as follows:

Years of life (X):	1	2	3	4	5
Profits (Y) (in lakh of Rs.):	1250	1400	1650	1950	2300

Estimate the profit of a company in the 6th year. [Ans. $Y = 265X + 915$, $Y_6 = $ Rs.2505 lakh]

9. From the following data:
(*i*) Obtain the two regression equations.
(*ii*) Draw up two regression lines on the graph paper.

X:	65	66	67	68	69	70	71
Y:	67	68	64	70	70	69	68

[Ans. $X = 0.462Y + 36.72$, $Y = 0.353X + 44$]

(ii) Using Deviations taken from Assumed Means

When actual means turn out to be in fractions rather than the whole numbers like 24.69, 25.12 etc., then it becomes difficult to take deviations from actual means and squaring them up. To avoid such difficulty, deviations from assumed means rather than actual means are used. In such case, regression equations are expressed as follows:

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

Here, byx = Regression coefficient of Y on X.

Using deviations from assumed means, the value of byx can be calculated as:

$$byx = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dx^2 - (\Sigma dx)^2}$$

or

$$byx = \frac{\Sigma dxdy - \dfrac{\Sigma dx \,.\, \Sigma dy}{N}}{\Sigma dx^2 - \dfrac{(\Sigma dx)^2}{N}}$$

Where, $dx = X - A_x$, $dy = Y - A_y$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

Where, bxy = Regression coefficient of X on Y.

Using deviations from assumed means, the value of bxy can be calculated as:

$$bxy = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}$$

or

$$bxy = \frac{\Sigma dxdy - \dfrac{\Sigma dx \,.\, \Sigma dy}{N}}{\Sigma dy^2 - \dfrac{(\Sigma dy)^2}{N}}$$

Where, $dx = X - A_x$, $dy = Y - A_y$

The following examples will clarify this method.

Example 16. Obtain the two regression equations for the following data:

X:	43	44	46	40	44	42	45	42	38	40	52	57
Y:	29	31	19	18	19	27	27	29	41	30	26	10

Also find the value of X when $Y = 49$ and Y when $X = 50$. Hence or otherwise find 'r'.

Solution:

Calculation of Regression Equations

X	A = 42 dx	dx^2	Y	A = 27 dy	dy^2	$dxdy$
43	1	1	29	2	4	2
44	2	4	31	4	16	8
46	4	16	19	−8	64	−32
40	−2	4	18	−9	81	18
44	2	4	19	−8	64	−16
42	0	0	27 = A	0	0	0
45	3	9	27	0	0	0
42 = A	0	0	29	2	4	0
38	−4	16	41	14	196	−56
40	−2	4	30	3	9	−6
52	10	100	26	−1	1	−10
57	15	225	10	−17	289	−255
$N = 12$ $\Sigma X = 533$	$\Sigma dx = 29$	$\Sigma dx^2 = 383$	$\Sigma Y = 306$ $N = 12$	$\Sigma dy = -18$	$\Sigma dy^2 = 728$	$\Sigma dxdy = -347$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{533}{12} = 44.42, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{306}{12} = 25.5$$

Since the actual means of X and Y are in fractions, we should take deviations from assumed mean to simplify the calculations.

$$byx = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{12 \times (-347) - (29)(-18)}{12 \times 383 - (29)^2} = \frac{-4164 + 522}{4596 - 841} = \frac{-3642}{3755}$$

$$= -0.969 = -0.97$$

$$bxy = \frac{N \times \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dy^2 - (\Sigma dy)^2}$$

$$= \frac{12 \times (-347) - (29)(-18)}{12 \times 728 - (-18)^2} = = \frac{-4164 + 522}{8736 - 324} = \frac{-3642}{8412}$$

$$= -0.432 = -0.43$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

$$X - 44.42 = -0.43\,(Y - 25.5)$$

$$X - 44.42 = -0.43Y + 10.965$$

$$X = -0.43Y + 55.385$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 25.5 = -0.97\,(X - 44.42)$$

$$Y - 25.5 = -0.97X + 43.0874$$

$$Y = -0.97X + 68.5874$$

When $Y = 49$,

$$X = -0.43Y + 55.385$$

$$= -0.43\,(49) + 55.385$$

$$= -21.07 + 55.385$$

$$\therefore \quad X_{49} = 34.315$$

When $X = 50$,

$$Y = -0.97(50) + 68.5874$$

$$= -48.5 + 68.5874$$

$$= 20.0874$$

$$\therefore \quad Y_{50} = 20.0874$$

Coefficient of Correlation

$$r = \sqrt{byx \,.\, bxy}$$

$$= -\sqrt{(-0.97) \times (-0.43)} = -0.645$$

Example 17. Obtain the regression equation of Y on X from the following data:

X:	78	89	97	69	59	79	68	61
Y:	125	137	156	112	107	136	124	108

Solution:

Calculation of Regression Equations

X	A = 69 dx	dx^2	Y	A = 112 dy	dy^2	$dxdy$
78	+9	81	125	+13	169	+117
89	+20	400	137	+25	625	+500
97	+28	784	156	+44	1936	+1232
69 = A	0	0	112 = A	0	0	0
59	−10	100	107	−5	25	+50
79	+10	100	136	+24	576	+240
68	−1	1	124	+12	144	−12
61	−8	64	108	−4	16	+32
$N = 8$ $\Sigma X = 600$	$\Sigma dx = 48$	$\Sigma dx^2 = 1530$	$\Sigma Y = 1005$	$\Sigma dy = 109$	$\Sigma dy^2 = 3491$	$\Sigma dxdy = 2159$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{600}{8} = 75, \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{1005}{8} = 125.625$$

$$byx = \frac{N \,.\, \Sigma dxdy - \Sigma dx \,.\, \Sigma dy}{N \,.\, \Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{8 \times 2159 - (48)(109)}{8 \times 1530 - (48)^2} = \frac{17272 - 5232}{12240 - 2304} = \frac{12040}{9936} = 1.212$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 125.625 = 1.212\,(X - 75)$$

$$Y - 125.625 = 1.212X - 90.9$$

$$Y = 1.212X + 34.725$$

IMPORTANT TYPICAL EXAMPLES

Example 18. A panel of judges A and B graded seven independently and awarded the following marks

Debater:	1	2	3	4	5	6	7
Marks by A:	40	34	28	30	44	38	?
Marks by B:	32	39	26	30	38	34	31

An eight debator was awarded 36 marks by Judge A while Judge B was not present the Judge B was also present, how many marks would you expect him to award eighth debator assuming degree of relationship exists in judgement?

Solution: Let marks awarded by Judge A be denoted by X and marks awarded by judge B denoted by Y. The marks expected to be awarded by Judge B can be determined by fitting regression equations of Y on X.

Calculation of Regression Equations

X	A = 30 dx	dx^2	Y	A = 30 dy	dy^2	$dxdy$
40	+10	100	32	2	4	20
34	+4	16	39	9	81	36
28	−2	4	26	−4	16	8
30 = A	0	0	30 = A	0	0	0
44	14	196	38	8	64	112
38	8	64	34	4	16	32
31	1	1	28	−2	4	−2
$N = 7$ $\Sigma X = 245$	$\Sigma dx = 35$	$\Sigma dx^2 = 381$	$\Sigma Y = 227$	$\Sigma dy = 17$	$\Sigma dy^2 = 185$	$\Sigma dxdy = 246$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{245}{7} = 35, \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{227}{7} = 32.43$$

$$byx = \frac{N \cdot \Sigma dxdy - \Sigma dx \cdot \Sigma dy}{N \cdot \Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{7 \times 206 - (35)(17)}{7 \times 381 - (35)^2}$$

$$= \frac{1442 - 595}{2667 - 1225} = \frac{847}{1442} = 0.587$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 32.43 = 0.587\,(X - 35)$$

$$Y - 32.43 = 0.587X - 20.545$$

$$\therefore \quad Y = 0.587X + 11.885$$

For $X = 36$, Y shall be

$Y = 0.587\,(36) + 11.885 = 21.132 + 11.885 = 33.017$ or 33 approx.

Thus, if the Judge B was also present, he would have awarded 33 marks to the eighth debator.

Example 19. Simple observations obtained to study the relation between the measure of the waist and the length of the trousers are shown as under :

Measure of the Waist (in cm) :	70	72.5	75	77.5	80	82.5	85	87.5	90	92.5
Length of Trousers (in cm):	100	102	100	95	105	110	95	98	100	105

Obtain the line of best fit (regression) of length of trousers on measurement of the waist. Calculate the coefficient of determination.

Solution: Let X = measure of waist and Y = length of trousers.

Here, $N = 10$, $\Sigma X = 812.5$, $\Sigma Y = 1010$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{812.5}{10} = 81.25 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{1010}{10} = 101$$

Since, $\bar{X}$ is not an integer, we will take the deviation of X from assumed value. Taking $dx = X - 80$ and $dy = Y - 101$.

The calculations are:

X	dx	dx^2	Y	dy	dy^2	$dxdy$
70	−10	100	100	−1	1	10
72.5	−7.5	56.25	102	+1	1	−7.5
75	−5	25	100	−1	1	+5
77.5	−2.5	6.25	95	−6	36	+15
80 = A	0	0	105	+4	16	0
82.5	25	6.25	110	+9	81	+22.5
85	5	25	95	−6	36	−30
87.5	7.5	56.25	98	−3	9	−22.5
90	10	100	100	−1	1	−10
92.5	12.5	156.25	105	+4	16	+50
$\Sigma X = 812.5$	$\Sigma dx = 12.5$	$\Sigma dx^2 = 531.25$	$\Sigma Y = 1010$	$\Sigma dy = 0$	$\Sigma dy^2 = 198$	$\Sigma dxdy = 32.5$

Regression Coefficient of Y on X:

$$byx = \frac{N \times \Sigma dxdy - \Sigma dx \cdot \Sigma dy}{N \cdot \Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{(10 \times 32.5) - (12.5 \times 0)}{(10 \times 531.25) - (12.5)^2} = \frac{325}{5312.5 - 156.25}$$

$$= \frac{325}{5156.25} = 0.06$$

Line of regression of length of trousers on the measurement of the waist, i.e., line of regression of Y on X is

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 101 = 0.06\,(X - 81.25)$$
$$Y - 101 = 0.06X - 4.875$$
$$Y = 0.06X + 96.125$$

Coefficient of Determination:

$$r^2 = \left[\frac{N\,\Sigma dxdy - (\Sigma dx)\,.\,(\Sigma dy)}{\sqrt{N\,\Sigma dx^2 - (\Sigma dx)^2}\,\sqrt{N\,\Sigma dy^2 - (\Sigma dy)^2}}\right]^2$$

$$= \left[\frac{(10\times 32.5) - (12.5\times 0)}{\sqrt{10\times 531.25 - (12.5)^2}\,\sqrt{10\times 198 - 0}}\right]^2$$

$$= \left[\frac{325}{\sqrt{5312.5 - 156.25}\times\sqrt{1980}}\right]^2 = \left[\frac{325}{\sqrt{5156.25}\times\sqrt{1980}}\right]^2$$

$$= \frac{(325)^2}{5156.25\times 1980} = \frac{105625}{10209375} = 0.01$$

Example 20. For a bivariate data, you are given the following information:

$\Sigma(X-58) = 46$ $\quad \Sigma(X-58)^2 = 3086$

$\Sigma(Y-58) = 9$ $\quad \Sigma(Y-58)^2 = 483$

$\Sigma(X-58)(Y-58) = 1095$

$N = 7$

(Assumed means of X and Y series are both 58)

You are required to determine (*i*) the two regression equations and (*ii*) the coefficient of correlation between X and Y series.

Solution: Since the assumed means of X and Y series are both 58, we have,

$\Sigma dx = 46$, $\quad \Sigma dx^2 = 3086$

$\Sigma dy = 9$ $\quad \Sigma dy^2 = 483$

$\Sigma dxdy = 1095$ $\quad N = 7$

$$byx = \frac{N\,.\,\Sigma dxdy - \Sigma dx\,.\,\Sigma dy}{N\,.\,\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{7\times 1095 - (46)(9)}{7\times 3086 - (46)^2}$$

$$= \frac{7665 - 414}{21602 - 2116} = \frac{7251}{19486} = 0.37$$

$$bxy = \frac{N\,.\,\Sigma dxdy - \Sigma dx\,.\,\Sigma dy}{N\,.\,\Sigma dy^2 - (\Sigma dy)^2}$$

$$= \frac{7\times 1095 - (46)(9)}{7\times 483 - (9)^2} = \frac{7251}{3300} = 2.20$$

Further,

$$\bar{X} = A + \frac{\Sigma\,dx}{N} = 58 + \frac{46}{7} = 64.57$$

$$\bar{Y} = A + \frac{\Sigma\,dy}{N} = 58 + \frac{9}{7} = 59.29$$

Regression Equation of X on Y:

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 64.57 = 2.20\,(Y - 59.29)$$
$$X - 64.57 = 2.20Y - 130.44$$
$$X = 2.20Y - 130.44 + 64.57$$
$$X = 2.20Y - 65.87$$

Regression Equation of Y on X:

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 59.29 = 0.37(X - 64.57)$$
$$Y - 59.29 = 0.37X - 23.891$$
$$Y = 0.37X - 23.891 + 59.29$$
$$Y = 0.37X + 35.399$$

$$r = \sqrt{bxy\cdot byx}$$

$$r = \sqrt{2.20\times 0.37} = 0.902$$

EXERCISE 2.4

1. Obtain the two regression equations for the following data:

X:	8	6	4	7	5	3
Y:	9	8	5	6	2	6

Also find the coefficient of correlation from the regression coefficients.

[**Ans.** $X = 3.1 + 0.4Y$; $Y = 2.23 + 0.685X$; $r = 0.523$]

2. Obtain the two regression equations from the following data:

Age of husband (X):	18	19	20	21	22	23	24	25	26	27
Age of wife (Y):	17	17	18	18	18	19	19	20	21	21

Also find the coefficient of correlation from the regression coefficients.

[**Ans.** $Y = 0.47X + 8.225$, $X = 1.99Y - 14.9$, $r = +0.967$]

3. The height of fathers and sons in inches are:

Height of Fathers:	65	66	68	69	71	73	67	68	70	72	65
Height of Sons:	67	68	64	72	70	69	70	68	68	73	67

Estimate (*i*) the height of son if the height of the father is 64 inches, and (*ii*) the height of father if the height of son is 71.
Also calculate the value of Spearman's coefficient of correlation between them.
[Ans. (*i*) 66.18, (*ii*) 69.2, (*iii*) $R_s = 0.4545$]

4. The age and blood pressure of 10 university teachers are:

Age :	56	42	36	47	49	42	60	72	63	57
Blood Pressure:	147	125	118	128	145	140	155	160	149	150

(*i*) Find the correlation coefficient between age and blood pressure.
(*ii*) Determine the least square regression equation of blood pressure on age.
(*iii*) Estimate the blood pressure of a teacher whose age is 45 years.
[**Hint:** See Example 51] [Ans. $r = 0.89$, $Y = 1.11X + 83.758$, $Y_{45} = 133.708 \approx 134$]

5. The following table gives age (X) in years of cars and annual maintenance cost (Y) in hundred rupees:

X:	1	3	5	7	9
Y:	15	18	21	23	22

Estimate the maintenance cost for a 4 year old car after finding the regression equation.
[Ans. $Y = 0.95X + 15.05$; $Y_4 = 18.85$]

6. Obtain the two regression equations from the following data:

X:	4	5	6	8	11
Y:	12	10	8	7	5

Verify that the coefficient of correlation is the geometric mean of the two regression coefficients.
[**Hint:** See Example 50]
[Ans. $X = 15.024 - 0.979Y$; $Y = -0.929X + 14.717$; $r = -0.956$]

7. Calculate from the following data:
(*i*) Two regression equations
(*ii*) Coefficient of correlation
(*iii*) Most likely value of X when Y = 10.

X:	45	55	56	58	60	65	68	70	75	80	85
Y:	56	50	48	60	62	64	65	70	74	82	90

[Ans. $X = 9.403 + 0.8514Y$, $Y = 0.884 + 0.992X$, $r = 0.9187$, $X_{10} = 17.915$]

(4) To Obtain Regression Equations from Coefficient of Correlation, Standard Deviations and Arithmetic Means of X and Y:
When the values of $\bar{X}$ and $\bar{Y}$, σ_x and σ_y and r of X and Y series are given, then regression equations are expressed in the following manner:

(1) Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X}) \qquad \text{where, } byx = r.\frac{\sigma_y}{\sigma_x}$$

or $$Y - \bar{Y} = r.\frac{\sigma_y}{\sigma_x}(X - \bar{X})$$

(2) Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y}) \qquad \text{where, } bxy = r.\frac{\sigma_x}{\sigma_y}$$

or $$X - \bar{X} = r.\frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$

Note: The above said form, of the regression equation is used only when the values of $\bar{X}$ and $\bar{Y}$, σ_x and σ_y and r are given.
The following examples makes the above said method more clear.

Example 21. You are given the following information:

	X	Y
Arithmetic mean:	5	12
Standard deviation:	2.6	3.6
Correlation coefficient:	$r = 0.7$	

(*i*) Obtain two regression equations.
(*ii*) Estimate Y when X = 9.
(*iii*) Estimate X when Y = 12.

Solution: Given, $\bar{X} = 5, \bar{Y} = 12, \sigma_x = 2.6, \sigma_y = 3.6, r = 0.7$

(*i*) Regression Equation of X on Y

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

Putting the values in the equation, we get

$$X - 5 = 0.7 \times \frac{2.6}{3.6}(Y - 12)$$

$$X - 5 = 0.51\,(Y - 12)$$

$$X - 5 = 0.51Y - 6.12$$

$$X = 0.51Y - 1.12$$

or $$X = -1.12 + 0.51Y$$

Regression Equation of Y on X

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

Putting the values in the equation, we get

$$Y - 12 = 0.7 \times \frac{3.6}{2.6}(X - 5)$$

$$Y - 12 = 0.97(X - 5)$$
$$Y - 12 = 0.97X - 0.97 \times 5$$
$$Y - 12 = 0.97X - 4.85$$
$$Y = 0.97X - 4.85 + 12$$
$$Y = 0.97X + 7.15$$

or $Y = 7.15 + 0.97X$

(ii) Most likely value of Y when X = 9

For this purpose, we use regression of Y on X

$$Y = 7.15 + 0.97X$$

Putting X = 9 in the equation, we get

$$Y = 7.15 + 0.97(9) = 7.15 + 8.73 = 15.88$$

(iii) Most likely value of X when Y = 12

For this purpose, we use regression of X on Y

$$X = -1.12 + 0.51Y$$

Putting Y = 12 in the equation, we get

$$X = -1.12 + 0.51(12)$$
$$X = -1.12 + 6.12 = 5$$

Example 22. You are given below the following information about advertisement and sales:

	Adv. Expenditure (Rs. crore)	Sales (Rs. crore)
Mean	20	120
S.D.	5	25
Correlation coefficient, $r_{xy} = +0.8$		

(*i*) Calculate the two regression equations.

(*ii*) What should be the advertisement budget if the company wants to attain sale target of Rs. 150 crore?

(*iii*) Find the most likely sales when advertisement expenditure is Rs. 25 crore.

Solution: Let X = Adv. Expenditure and Y = Sales

Thus, we have $\bar{X} = 20, \bar{Y} = 120, \sigma_x = 5, \sigma_y = 25, r_{xy} = 0.8$

(*i*) **(a) Regression Equation of X on Y**

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

$$X - 20 = 0.8 \times \frac{5}{25}(Y - 120)$$

$$X - 20 = 0.16(Y - 120)$$
$$X - 20 = 0.16Y - 19.2$$
$$X = 0.16Y - 19.2 + 20$$

$\therefore$ $X = 0.16Y + 0.8$

(b) Regression Equation of Y on X

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

$$Y - 120 = 0.8 \times \frac{25}{5}(X - 20)$$

$$Y - 120 = 4(X - 20)$$
$$Y - 120 = 4X - 80$$
$$Y = 40 + 4X$$

(*ii*) When sales target (Y) is Rs. 150 crore, then the advertisement expenditure (X) is

$$X = 0.8 + 0.16Y$$

Put $Y = 150$, $X = 0.8 + 0.16(150)$

$= 0.8 + 24 = 24.8$ crore.

(*iii*) When advertisement expenditure (X) is Rs. 25 crore, the sales (Y) is

$$Y = 40 + 4X$$

Put $X = 25$, $Y = 40 + 4(25)$

$= 40 + 100 = 140$ crore.

Example 23. Find the regression equations when you know:

$$\bar{X} = 68.2,\ \bar{Y} = 9.9,\ \frac{\sigma_y}{\sigma_x} = 0.44, r = 0.76$$

Solution: Given, $\bar{X} = 68.2,\ \bar{Y} = 9.9,\ \frac{\sigma_y}{\sigma_x} = 0.44, r = 0.76$

(*i*) **Regression Equation of Y on X**

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

Putting the values in the equation, we get

$$Y - 9.9 = 0.76 \times 0.44 (X - 68.2)$$
$$Y - 9.9 = 0.3344 (X - 68.2)$$
$$Y - 9.9 = 0.3344X - 22.81$$
$$Y = 0.3344X - 22.81 + 9.9$$
or $$Y = 0.3344X - 12.91$$

(ii) Regression Equation of X on Y:

When $\frac{\sigma_y}{\sigma_x} = 0.44$ or $\frac{44}{100}$

Then $\frac{\sigma_x}{\sigma_y} = \frac{100}{44}$ or 2.27

$$X - \bar{X} = r \cdot \frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$
$$X - 68.2 = 0.76 \times 2.27 (Y - 9.9)$$
$$X - 68.2 = 1.725 (Y - 9.9)$$
$$X - 68.2 = 1.725Y - 17.08$$
$$X = 1.725Y - 17.08 + 68.2$$
$$X = 1.725Y + 51.12$$

Example 24. Find the expected price in Mumbai when price in Calcutta is Rs. 70 using the following data:

Average Price in Calcutta	:	Rs. 65
Average Price in Mumbai	:	Rs. 67
S.D. of Price in Calcutta	:	2.5
S.D. of Price in Mumbai	:	3.5
Correlation coefficient between price of Mumbai and Calcutta	:	0.8

Solution: Let X = Price in Calcutta, Y = Price in Mumbai

Given: $\bar{X} = 65, \bar{Y} = 67, \sigma_x = 2.5, \sigma_y = 3.5, r = 0.8$

Expected Price in Mumbai (Y) when price in Calcutta (X) = 70 can be found from regression equation of Y on X.

Regression Equation of Y on X

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

Putting the values, we get

$$Y - 67 = 0.8 \times \frac{3.5}{2.5}(X - 65)$$
$$Y - 67 = 1.12 (X - 65)$$

$$Y - 67 = 1.12X - 72.8$$
$$Y = 1.12X - 72.8 + 67$$
$$Y = 1.12X - 5.8$$

When $X = 70$, $Y = 1.12 (70) - 5.8 = 78.4 - 5.8$
$= 72.6$

Thus, the expected price in Mumbai is Rs. 72.6 corresponding to Rs. 70 at Calcutta.

Example 25. The coefficient of correlation between the ages of husbands and wives in a community was found to be +0.8, the average of husband age was 25 years and that of wives age 22 years. Their standard deviations were 4 and 5 years respectively. Find with the help of regression equations:

(i) the expected age of husband when wife's age is 20 years and

(ii) the expected age of wife when husband's age is 33 years.

Solution: Let age of wife be denoted by Y and age of husband by X. We are given:

$\bar{X} = 25, \bar{Y} = 22, \sigma_x = 4, \sigma_y = 5, r = 0.8$

(i) For estimating age of husband when wife's age is 20 years, we use regression of X on Y as follows:

$$X - \bar{X} = r \cdot \frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$
$$X - 25 = 0.8 \times \frac{4}{5}(Y - 22)$$
$$X - 25 = 0.64 (Y - 22)$$
$$X - 25 = 0.64Y - 14.08$$
$$X = 0.64Y + 10.92$$

When Y = 20, $X = 0.64 (20) + 10.92 = 12.8 + 10.92 = 23.72$

Thus, the expected age of husband when wife's age is 20 years shall be 23.72 years.

(ii) For estimating age of wife when husband's age is 33 years, we use regression equation of Y on X as follows:

$$Y - \bar{Y} = r \cdot \frac{\sigma_y}{\sigma_x}(X - \bar{X})$$
$$Y - 22 = 0.8 \times \frac{5}{4}(X - 25)$$
$$Y - 22 = 1 (X - 25)$$
$$Y - 22 = X - 25 \quad \Rightarrow \quad Y = X - 3$$

When X = 33, Y = 33 − 3 = 30

Thus, the expected age of wife when husband's age is 33 is 30 years.

IMPORTANT TYPICAL EXAMPLES

Example 26. The following data based on 450 students are given for marks in Statistics and Economics at a certain Examination:

Mean Marks in Statistics	:	40
Mean Marks in Economics	:	48
S.D. of Marks in Statistics	:	12
The variance of marks in Economics	:	256

Sum of the products of deviations of marks from their respective means is 42075.

(*i*) Obtain the equations of two lines of regression.

(*ii*) Estimate the average marks in Economics of candidates who obtained 50 marks in Statistics.

Solution: (*i*) Let X denote marks in Statistics and Y denote marks in Economics. We are given:

$\bar{X} = 40$, $\bar{Y} = 48$

$\sigma_x = 12$, $\sigma_y^2 = 256 \Rightarrow \sigma_y = 16$

$\Sigma xy = 42075$

Before we obtain the regression equations, we compute the coefficient of correlation (r) by using the formula:

$$r = \frac{\Sigma xy}{N.\sigma_x.\sigma_y}$$

$$= \frac{42075}{450 \times 12 \times 16} = \frac{42075}{86400}$$

$= +0.49$ approx.

Regression Equation of X on Y

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

$$X - 40 = 0.49 \times \frac{12}{16}(Y - 48)$$

$$X - 40 = \frac{5.88}{16}(Y - 48)$$

$$X - 40 = 0.3675(Y - 48)$$

$$X - 40 = 0.3675Y - 17.64$$

$$X = 0.3675Y - 17.64 + 40$$

$$X = 0.3675Y + 22.36$$

Regression Equation of Y on X

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

$$Y - 48 = 0.49 \times \frac{16}{12}(X - 40)$$

$$Y - 48 = \frac{7.84}{12}(X - 40)$$

$$Y - 48 = 0.653(X - 40)$$

$$Y - 48 = 0.653X - 26.12$$

$$Y = 0.653X - 26.12 + 48$$

$$Y = 0.653X + 21.88$$

(*ii*) To estimate the marks in Economics when 50 marks in Statistics is given, we use regression of Y on X.

$$Y = 0.653X + 21.88$$

When X = 50,

$$Y = 0.653\ (50) + 21.88$$

$$= 32.65 + 21.88$$

$$= 54.53 \text{ or } 55 \text{ marks}$$

Thus, the expected marks in Economics is 55.

Example 27. If $\bar{X} = 25$, $\bar{Y} = 120, bxy = 2$

Estimate the value of X when Y = 130.

Solution: Given, $\bar{X} = 25$, $\bar{Y} = 120, bxy = 2$

For estimating X when $Y = 130$, we use regression equation of X on Y as follows:

$$X - \bar{X} = bxy(Y - \bar{Y})$$

or

$$X = \bar{X} + bxy(Y - \bar{Y})$$

$$X = 25 + 2(130 - 120)$$

$$X = 25 + 2(10) = 45$$

Thus, the value of X is 45 when $Y = 130$.

Example 28. If $\sigma_x^2 = 9, \sigma_y^2 = 1600, r_{xy} = 0.5$, obtain bxy.

Solution: Given, $\sigma_x^2 = 9 (or\ \sigma_x = 3), \sigma_y^2 = 1600 (or\ \sigma_y = 40), r_{xy} = 0.5$,

We know, $bxy = r.\frac{\sigma_x}{\sigma_y}$

$\therefore$ $bxy = 0.5 \times \frac{3}{40} = \frac{1.5}{40}$

$= 0.0375$

Example 29. The regression coefficient of Y on X, *i.e.*, $byx = 1.2$. If $u = \frac{X-100}{2}$ and $v = \frac{Y-100}{3}$, Find bvu.

Solution: $byx = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\Sigma(X-\bar{X})^2}$

$u = \frac{X-100}{2}$ $\qquad$ $v = \frac{Y-100}{3}$

$\Rightarrow X = 2u+100$ $\qquad$ $\Rightarrow Y = 3v+100$

$\bar{X} = 2\bar{u}+100$ $\qquad$ $\bar{Y} = 3\bar{v}+100$

$\therefore (X-\bar{X}) = 2(u-\bar{u})$ $\qquad$ $\therefore (Y-\bar{Y}) = 3(v-\bar{v})$

Now, $byx = \frac{2\times 3\Sigma(u-\bar{u})(v-\bar{v})}{4\cdot\Sigma(u-\bar{u})^2} = 1.5\, bvu$

So, $bvu = \frac{byx}{1.5}$

$bvu = \frac{1.2}{1.5} = 0.8$

EXERCISE 2.5

1. The following data relate to marks in English and Maths:

Mean marks in English	39.5
Mean marks in Maths	47.6
S.D. of marks in English	10.8
S.D. of marks in Maths	16.9
Coefficient of correlation between English and Maths = + 0.42	

(*i*) Obtain the two regression equations.

(*ii*) Calculate the expected average marks in Maths of candidates who received 50 marks in English. [**Ans.** $X = 0.268Y + 26.73$, $Y = 0.657X + 21.64$, 54.5]

2. There are two series of index numbers, P for price index and S for stock commodities. The mean and standard deviation of P are 100 and 8 respectively and S are 103 and 4. The correlation coefficient between the two series is 0.4. With this data work out a linear equation to read off values of P for various values of S. Can the same equation be used to read off the values of S for different values of P? If not, give the appropriate equation.

[**Ans.** $P = 0.8S + 17.6$; No; $S = 0.2P + 83$]

3. The following results were worked out from marks in Statistics and Mathematics in a certain examination:

	Marks in Statistics (X)	Marks in Maths (Y)
A.M.	39.5	47.5
S.D.	10.8	17.8
Coefficient of correlation = 0.42		

(*i*) Find the two regression equations.

(*ii*) Estimate the value of Y when $X = 50$ and X, when $Y = 30$.

[**Ans.** (*i*) $Y = 0.6922X + 20.1581$, $X = 0.2548Y + 27.397$; (*ii*) 54.76, 35.04]

4. You are given the following data about sales and advertisement expenditure of a firm:

	Sales (Rs. crore)	Adv. Expenditure (Rs. crore)
Arithmetic Mean:	50	10
Standard Deviation:	10	2
Coefficient of correlation = +0.9		

(*i*) Calculate the two regression equations.

(*ii*) Estimate the likely sales for a proposed advertisement expenditure of Rs. 13.5 crore.

(*iii*) What should be the advertisement budget if the company wants to achieve a sales target of Rs. 70 crore?

[**Ans.** (*i*) $X = 4.5Y + 5$, $Y = 0.18X + 1$, (*ii*) 65.75 crores, (*iii*) 13.6 crores]

5. Given the following data, what would be the possible yield when rainfall is 29"?

	Rainfall	Yield per acre
Mean	25"	40
Variance	9"	36
Coefficient of correlation between rainfall and production = 0.8		

[**Ans.** 46.4]

6. For a given set of bivariate data, the following results were obtained:
$\bar{X} = 53.2$, $\bar{Y} = 27.9$, regression coefficient of Y on $X = -1.5$
Regression coefficient of X on $Y = -0.2$. Find the most probable value of Y when $X = 60$.
Also find the coefficient of correlation.

[**Ans.** $Y_{60} = 17.7$, $r = -0.548$]

7. If $\bar{X} = 45$, $\sigma_x = 2.5$, $\bar{Y} = 60$, $\sigma_y = 2.2$, $r = 0.75$
Estimate (*i*) Value of Y when $X = 35$
(*ii*) Value of X when $Y = 20$.

[**Ans.** (*i*) 53.4, (*ii*) 10.92]

8. Following information are given:

	X Series	Y Series
Mean:	10	8
Standard Deviation:	6	5

The covariance between X and Y is 15. Estimate the value of X when $Y = 9$.
[Ans. $X = 0.6Y + 5.2, X_{Y=9} = 10.6$]

9. If $\bar{Y} = 15, \bar{X} = 3.5, b_{yx} = 2.5$
Obtain estimate of Y when $X = 5$.

10. If $\sigma_x^2 = 0.75, \sigma_y^2 = 1.2, r_{xy} = 0.65$, find b_{xy}. [Ans. $b_{xy} = 0.52$]

11. If $\sigma_x^2 = 25, \sigma_y^2 = 625, bxy = 0.16$, find '$r$'. [Ans. $r = 0.8$]

12. If $byx = 0.50, bxy = 1.5$, find r. [Ans. $r = 0.87$]

13. A group of 20 students was observed for weight (X) and height (Y). The variance of height was found to be 9 cm and that of weight 1600 gm. If the correlation coefficient between height and weight was 0.5, obtain an average absolute increase in weight in response to height.
[Hint: Find bxy] [Ans. 6.67 gm]

14. In a regression analysis, the following two regression coefficients were obtained
$bxy = 3.5$ and $byx = 0.5$
Comment on these values. [Ans. $r^2 = 1.75 \Rightarrow r = 1.32 > 1$, Inconsistent value]

◘ TO OBTAIN REGRESSION EQUATIONS IN CASE OF GROUPED DATA

For obtaining regression equations from grouped data, first of all we have to construct a correlation table. After that, we find out $\bar{X}, \bar{Y}$ and the regression coefficients byx and bxy. Special adjustment must be made while calculating the value of regression coefficients because regression coefficients are independent of change of origin but not of scale. In grouped data, the regression coefficients (byx and bxy) are computed by using the following formulae:

(*i*) $$bxy = \frac{N \times \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{N \times \Sigma fdy^2 - (\Sigma fdy)^2} \times \frac{i_x}{i_y}$$

(*ii*) $$byx = \frac{N \,.\, \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{N \,.\, \Sigma fdx^2 - (\Sigma fdx)^2} \times \frac{i_y}{i_x}$$

Where, i_x = Common factor of X–variable
i_y = Common factor of Y–variable.

The following examples makes the computation of regression equations more clear:

Example 30. Obtain the two regression equations from the following bivariate frequency distribution:

Y \ X	0—20	20—40	40—60
10—25	10	5	3
25—40	4	40	8
40—55	6	9	15

Also compute Karl Pearsons' coefficient correlation from two regression coefficients.
(Table Given at Page 120)

Solution:

$$\bar{X} = A + \frac{\Sigma fdx}{N} \times i_x$$

$$= 30 + \frac{6}{100} \times 20$$

$$= 30 + \frac{120}{100}$$

$$= 30 + 1.2 = 31.2$$

$$\bar{Y} = A + \frac{\Sigma fdy}{N} \times i_y$$

$$= 32.5 + \frac{12}{100} \times 15$$

$$= 32.5 + 1.8 = 34.3$$

$$bxy = \frac{N \times \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{N \times \Sigma fdy^2 - (\Sigma fdy)^2} \times \frac{i_x}{i_y}$$

$$= \frac{(100)(16) - (6)(12)}{(100)(48) - (12)^2} \times \frac{20}{15}$$

$$= \frac{1600 - 72}{4800 - 144} \times \frac{20}{15} = \frac{1528}{4656} \times \frac{20}{15}$$

$$= \frac{30560}{69840} = +0.43$$

$$byx = \frac{N \,.\, \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{N \,.\, \Sigma fdx^2 - (\Sigma fdx)^2} \times \frac{i_y}{i_x}$$

$$= \frac{(100)(16) - (6)(12)}{(100)(46) - (6)^2} \times \frac{15}{20}$$

$$= \frac{1600 - 72}{4600 - 36} \times \frac{15}{20} = \frac{1528}{4564} \times \frac{15}{20}$$

$$= \frac{22920}{91280} = +0.25$$

Solution 30.

Let $dx = \frac{X-30}{20}$, $dy = \frac{Y-32.5}{15}$

Y ↓ M.V.	X → M.V.	0—20	20—40	40—60	f	fdy	fdy^2	$fdxdy$
	dx	−1	0	+1				
	dy	10	30	50				
10–25	17.5	−20	0	+20				
		−1	0	+1				
10–25	−15	1 10 10	0 5 0	−1 3 −3	18	−18	18	7
25–40	32.5	0 4 0	0 40 0	0 8 0	52	0	0	0
40–55	47.5	−1 6 −6	0 9 0	1 15 15	30	+30	30	9
	f	20	54	26	$N = 100$	$\Sigma fdy = +12$	$\Sigma fdy^2 = 48$	$\Sigma fdxdy = 16$
	fdx	−20	0	+26	$\Sigma fdx = 6$			
	fdx^2	20	0	+26	$\Sigma fdx^2 = 46$			
	$fdxdy$	4	0	12	$\Sigma fdxdy = 16$			

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 34.3 = 0.25\,(X - 31.2)$$
$$Y - 34.3 = 0.25X - 7.8$$
$$Y = 0.25X + 26.5$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 31.2 = 0.43\,(Y - 34.3)$$
$$X - 31.2 = 0.43Y - 14.749$$
$$X = 0.43Y + 16.451$$

Coefficient of Correlation (r)

$$r = \sqrt{bxy \,.\, byx}$$
$$= \sqrt{(+0.43)(+0.25)}$$

$\therefore r = +0.33$ approx.

Alter: r can also be computed by using the formula:

$$r = \frac{N\,.\,\Sigma fdxdy - \Sigma fdx\,.\,\Sigma fdy}{\sqrt{N\,.\,\Sigma fdx^2 - (\Sigma fdx)^2}\ \sqrt{N\,.\,\Sigma fdy^2 - (\Sigma fdy)^2}}$$
$$= \frac{100 \times 16 - (6)(+12)}{\sqrt{100 \times 46 - (6)^2}\ \sqrt{100 \times 48 - (12)^2}}$$
$$= \frac{1600 - 72}{\sqrt{4600 - 36}\ \sqrt{4800 - 144}} = \frac{1528}{\sqrt{4564}\ \sqrt{4656}}$$
$$= \frac{1528}{4609.77} = 0.33$$

Example 31. Following is the distribution of students according to their height and weight:

Height (in inches)	Weight (in lbs.)			
	90—100	100—110	110—120	120—130
50—55	4	7	5	2
55—60	6	10	7	4
60—65	6	12	10	7
65—70	3	8	6	3

Calculate: (*i*) the two regression coefficients.
(*ii*) the two regression equations.
(*iii*) coefficient of determination.

Solution: **(Table given at page 122)** Let X denote height and Y denote weight.

Solution 31.

Let $dx = \frac{X - 57.5}{5}$, $dy = \frac{Y - 105}{10}$

X ↓ M.V. \ Y → M.V. ; dy \ dx		90—100	100—110	110—120	120—130	f	fdx	fdx^2	$fdxdy$
		95	105	115	125				
		−10	0	+10	+20				
		−1	0	+1	+2				
50—55	52.5	−5	1 4 4	0 7 0	−1 5 −5	−2 2 −4	18	−18	18	−5
55—60	57.5	0	0 6 0	0 10 0	0 7 0	0 4 0	27	0	0	0
60—65	62.5	+5	−1 6 −6	1 10 10	2 7 14	4 3 12	35	35	35	18
65—70	67.5	+10	−2 3 −6	2 6 12	4 3 12	0 8 0	20	40	80	18
		f	19	28	16	N=100	57	133	31
		fdy	−19	+28	32	41			
		fdy^2	19	28	64	111			
		$fdxdy$	−8	17	22	31			

(i) Regression Coefficient of Y on X

$$byx = \frac{N \cdot \Sigma fdxdy - \Sigma fdx \cdot \Sigma fdy}{N \cdot \Sigma fdx^2 - (\Sigma fdx)^2} \times \frac{i_y}{i_x}$$

$$byx = \frac{100 \times 31 - (57)(41)}{100 \times 133 - (57)^2} \times \frac{10}{5}$$

$$= \frac{3100 - 2337}{13300 - 3249} \times \frac{2}{1} = \frac{763}{10051} \times 2 = \frac{1526}{10051} \approx 0.151 \approx 0.15$$

(ii) Regression Coefficient of X on Y

$$bxy = \frac{N \times \Sigma fdxdy - \Sigma fdx \cdot \Sigma fdy}{N \times \Sigma fdy^2 - (\Sigma fdy)^2} \times \frac{i_x}{i_y}$$

$$= \frac{100 \times 31 - (57)(41)}{100 \times 111 - (41)^2} \times \frac{5}{10} = \frac{3100 - 2337}{11100 - 1681} \times \frac{1}{2}$$

$$= \frac{763}{9419} \times \frac{1}{2} = \frac{763}{18838} = 0.040 \approx 0.04$$

Calculation of $\bar{X}$ and $\bar{Y}$

$$\bar{X} = A + \frac{\Sigma fdx}{N} \times i_x$$

$$= 57.5 + \frac{57}{100} \times 5 = 57.5 + \frac{285}{100}$$

$$= 57.5 + 2.85 = 60.35$$

$$\bar{Y} = A + \frac{\Sigma fdy}{N} \times i_y$$

$$= 105 + \frac{41}{100} \times 10 = 105 + \frac{410}{100}$$

$$= 105 + 4.10 = 109.10$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 109.10 = 0.15\,(X - 60.35)$$

$$Y = 0.15X - 9.0525 + 109.10$$

$$Y = 0.15X + 100.04$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

$$X - 60.35 = 0.04\,(Y - 109.10)$$

$$X - 60.35 = 0.04Y - 4.364$$

$$X = 0.04Y + 55.98$$

(iii) Correlation Coefficient (r)

$$r = \sqrt{byx \,.\, bxy}$$

$$= \sqrt{0.15 \times 0.04} = \sqrt{0.006} = 0.077$$

Thus, Coefficient of Determination $= r^2 = (0.077)^2 = 0.006$

Example 32. From the following grouped data, find two regression equations and correlation co-efficient.

Wife's Age	Husband's Age			
	20—25	25—30	30—35	35—40
15—20	20	10	3	—
20—25	4	28	6	2
25—30	—	5	11	4
30—35	—	—	2	1
35—40	—	—	—	5

Solution: **(Table Given at Page 125)**

Regression Coefficient of X on Y

$$bxy = \frac{N \times \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{N \times \Sigma fdy^2 - (\Sigma fdy)^2} \times \frac{i_x}{i_y}$$

$$= \frac{100(138) - (-80)(-100)}{100(204) - (-100)^2} \times \frac{5}{5}$$

$$= \frac{13800 - 8000}{20400 - 10000} = \frac{5800}{10400} = 0.558$$

$\therefore$ $bxy = 0.558$

Regression Coefficient of Y on X

$$byx = \frac{N \,.\, \Sigma fdxdy - \Sigma fdx \,.\, \Sigma fdy}{N \,.\, \Sigma fdx^2 - (\Sigma fdx)^2} \times \frac{i_y}{i_x}$$

$$= \frac{100(138) - (-80)(-100)}{100(150) - (-80)^2} \times \frac{5}{5}$$

$$= \frac{13800 - 8000}{15000 - 6400} = \frac{5800}{8600} = 0.674$$

$\therefore$ $byx = 0.674$

$$\bar{X} = A + \frac{\Sigma fdx}{N} \times i_x = 32.5 + \left(\frac{-80}{100}\right)(5) = 28.5$$

$$\bar{Y} = A + \frac{\Sigma fdy}{N} \times i_y = 27.5 + \left(\frac{-100}{100}\right)(5) = 22.5$$

Solution 32. Let $dx = \frac{X - 32.5}{5}$, $dy = \frac{Y - 27.5}{5}$

X → / Y ↓	M.V.	dx	20—25	25—30	30—35	35—40	f	fdy	fdy²	fdxdy
			22.5	27.5	32.5	37.5				
		dy	−10	−5	0	5				
15—20	17.5	−2	4 20 80	2 10 20	0 3 0	−2 2 −4	35	−70	140	96
20—25	22.5	−1	2 4 8	1 28 28	0 6 0	−1 4 −4	42	−42	42	32
25—30	27.5	0	—	0 5 0	0 11 0	—	16	0	0	0
30—35	32.5	1	—	—	0 2 0	—	2	2	2	0
35—40	37.5	2	—	—	—	2 5 10	5	10	20	10
		f	24	43	22	11	N = 100	−100	204	138
		fdx	−48	−43	0	11	= −80			
		fdx²	96	43	0	11	= +150			
		fdxdy	88	48	0	2	= +138			

Regression Equation of X on Y

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$
$$X - 28.5 = 0.558(Y - 22.5)$$
$$X - 28.5 = 0.558Y - 12.555$$
$$X = 15.945 + 0.558Y$$

Regression Equation of Y on X

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$
$$Y - 22.5 = 0.674(X - 28.5)$$
$$Y - 22.5 = 0.674X - 19.209$$
$$Y = 0.674X - 19.209 + 22.5$$
$$Y = 3.291 + 0.674X$$

Now $r = \sqrt{b_{xy} \times b_{yx}}$

$= \sqrt{0.558 \times 0.674} = +0.613$

EXERCISE 2.6

1. Obtain two regression equations for the following grouped data:

Sales (Rs. '000) (Y)	Advertisement Exp. (Rs. '000) (X)			
	5—15	15—25	25—35	35—45
75—125	4	1	—	—
125—175	7	6	2	—
175—225	1	3	4	1
225—275	1	1	3	2
				4

Also find coefficient of correlation. [Ans. $X = 0.134Y - 1.45$, where X denotes Adv Exp $Y = 2.65X + 119.13$ where Y denotes sales, $r = +0.59$]

2. Obtain the regression equations from the following data:

Marks in English (X)	Marks in Statistics (Y)					Total
	10—20	20—30	30—40	40—50	50—60	
10—20	6	3	—	—	—	9
20—30	3	16	10	—	—	29
30—40	—	10	15	7	—	32
40—50	—	—	7	10	4	21
50—60	—	—	—	4	5	9
Total	9	29	32	21	9	100

Also find coefficient of determination. [Ans. $X = 6.77 + 0.802Y$; $Y = 6.77 + 0.802X$, $r^2 = 0.802$]

3. The following table shows the frequency distribution of 50 couples classified according to their ages.

Age of Wives (X)	Age of Husband (Y)			Total
	20—25	25—30	30—35	
16—20	9	14	—	23
20—24	6	11	3	20
24—28	—	—	7	7
Total	15	25	10	50

Estimate (*i*) the age of husband when wife's age is 20 years and (*ii*) the age of wife when husband's age is 30 years.

[Ans. $X = 0.47Y + 8.03$; $Y = 0.72X + 12.02$; $Y = 26.42$ years; $X = 22.13$ years]

4. Find regression equations from the following data:

Marks in Statistics	Marks in Economics				Total
	4—8	8—12	12—16	16—20	
8—14	11	6	2	1	20
14—20	5	12	15	8	40
20—26	—	2	3	15	20
Total	16	20	20	24	80

[Ans. $X = 0.67Y + 1.27$; $Y = 0.61X + 9.3$]

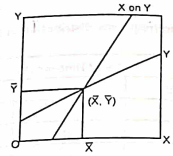
5. Find coefficient of correlation and regression equations from the following data:

Y \ X	5—15	15—25	25—35	35—45
0—10	1	1	—	—
10—20	3	6	5	1
20—30	1	8	9	2
30—40	—	3	9	3
40—50	—	—	4	4

[Ans. $r = +0.53$; $X = 15.58 + 0.42Y$; $Y = 8.91 + 0.67X$]

TO OBTAIN THE MEAN VALUES AND CORRELATION COEFFICIENT FROM THE REGRESSION EQUATIONS

(I) **To Find the Mean Values from the Regression Equations:** Two regression lines intersect each other at mean values ($\bar{X}$ and $\bar{Y}$) points. In other words, the point of intersection of the two lines of regression give the mean values of both X and Y variables, *i.e.*, $\bar{X}$ and $\bar{Y}$. This is clear from the following diagram:



The above diagram makes it clear that two regression lines Y on X and X on Y intersect each other at $\bar{X}$ and $\bar{Y}$ points. Therefore, it is clear that if both regression lines/regression equations are known, then solving them out gives the mean values X and Y series, *i.e.*, $\bar{X}$ and $\bar{Y}$.

(2) To Find the Coefficient of Correlation from two Regression Equations: Correlation coefficient can be worked out from the regression coefficients *bxy* and *byx*. From the regression equation of X on Y, we can find out *bxy* and from the regression equation of Y on X, we can find out *byx* and then correlation coefficient *r* can be derived as $r = \sqrt{byx \cdot bxy}$.

Note: But sometimes the regression equations are given in such way that by inspection, it is difficult to make out which one is the regression equation of X on Y and which one is Y on X. In such a case, any of them, taken as a regression equation of X on Y, is used to compute *bxy*. Similarly, with the help of other equation *byx* is computed. If the product of *bxy* and *byx* yields more than unity (1), then this follows that our supposition is wrong, because r^2 cannot exceed unity. That means we have to change our supposition. This time we have to make the supposition other way round, *i.e.*, previously taken to be X on Y should now be taken for Y on X. Thus, the product of regression coefficients shall not exceed unity and our results will be correct.

Alternative Method: To find out which regression equation is X on Y and which is Y on X, the following alternative method can also be applied:

Suppose two regression equations are as follows:

(1) $a_1x + b_1y + c_1 = 0$

(2) $a_2x + b_2y + c_2 = 0$

(i) If $a_1\ b_2 \leq a_2\ b_1$ (in magnitude, *i.e.*, ignoring signs), then

$a_1x + b_1y + c_1 = 0$ is the regression of Y on X and

$a_2x + b_2y + c_2 = 0$ is the regression of X on Y.

(ii) If $a_1\ b_2 > a_2\ b_1$ (in magnitude), then

$a_1x + b_1y + c_1 = 0$ is the regression of X on Y and

$a_2x + b_2y + c_2 = 0$ is the regression of Y on X.

Example 33. From the following two regression equations, identify which one is of X on Y and which one is of Y on X:

$$2X + 3Y = 42$$

$$X + 2Y = 26$$

Solution: In the absence of any clear cut indication, let us assume that equation first to be Y on X and equation second to be of X on Y.

Let equation first be regression equation of Y on X

$$2X + 3Y = 42$$

$$3Y = 42 - 2X$$

$$Y = \frac{42}{3} - \frac{2X}{3} \quad \text{...(i)}$$

$\Rightarrow$

From this it follows that

byx = Coefficient of X in $(i) = \frac{-2}{3}$

Now, equation (*ii*) be regression equation of X on Y

$$X + 2Y = 26$$

$$X = 26 - 2Y \quad \text{...(ii)}$$

From this it follows that

bxy = Coefficient of Y in $(ii) = -2$

Now, we calculate '*r*' on the basis of the above values of two regression coefficient, we get

$$r^2 = byx \cdot bxy = \frac{-2}{3} \times -2 = \frac{4}{3} > 1$$

Here, $r^2 > 1$ which is impossible as $r^2 \leq 1$. So, our assumption is wrong. We now choose equation (*i*) as regression of X on Y and (*ii*) as regression equation of Y on X:

Assuming the first equation as of X on Y, we have

$$2X + 3Y = 42$$

or

$$2X = 42 - 3Y$$

$$X = \frac{42}{2} - \frac{3Y}{2} \quad \text{...(iii)}$$

From this, it follows that

bxy = Coefficient of Y in $(iii) = \frac{-3}{2}$

Now, assuming the second equation as Y on X, we have

$$X + 2Y = 26$$

or

$$2Y = 26 - X$$

$\Rightarrow$

$$Y = \frac{26}{2} - \frac{1}{2}X \quad \text{...(iv)}$$

From this it follows that

byx = Coefficient of X in $(iv) = -\frac{1}{2}$

Now, $r^2 = bxy \cdot byx = -\frac{3}{2} \times \frac{-1}{2} = \frac{3}{4} = 0.75$

Here, $r^2 < 1$ which is possible. r^2 is within the limit, *i.e.*, $r^2 \leq 1$.

Hence, it is proved that the first equation is of X on Y and the second equation is of on X.

Aliter:

$$2X + 3Y = 42$$
$$X + 2Y = 26$$

As $2 \times 2 > 3 \times 1$

$\therefore$ $2X + 3Y = 42$ is the regression of X on Y and

$X + 2Y = 26$ is the regression of Y on X.

Example 34. Given the regression equations:

$$3X + 4Y = 44$$
$$5X + 8Y = 80$$

Variance of $X = 30$

Find $\bar{X}, \bar{Y}, r$ and σ_y

Solution: **Calculation of $\bar{X}$ and $\bar{Y}$**

The regression equations are:

$$3X + 4Y = 44$$
$$5X + 8Y = 80$$

Multiply (*i*) by 2 and substracting (*ii*) from it

$$6X + 8Y = 88$$
$$5X + 8Y = 80$$
$$- \quad - \quad -$$
$$X = 8 \quad \text{or} \quad \bar{X} = 8$$

Substituting the value of $X = 8$ in (*i*), we get

$$3(8) + 4Y = 44$$
$$24 + 4Y = 44$$
$$4Y = 20$$
$$Y = 5 \quad \text{or} \quad \bar{Y} = 5$$

$\therefore$ $\bar{X} = 8, \bar{Y} = 5$

Calculation of Correlation Coefficient

Suppose (*i*) be regression of Y on X

$$3X + 4Y = 44$$
$$4Y = 44 - 3X$$
$$Y = \frac{44}{4} - \frac{3}{4}X$$

$\therefore$ byx = Coefficient of X *in (iii)* $= -\frac{3}{4}$

Let equation (*ii*) be regression of X on Y

$$5X + 8Y = 80$$
$$5X = 80 - 8Y$$
$$X = \frac{80}{5} - \frac{8}{5}Y$$

...(*iv*)

$\therefore$ bxy = Coefficient of Y in (*iv*) $= -\frac{8}{5}$

Now $r^2 = byx \cdot bxy = -\frac{3}{4} \times \frac{-8}{5} = \frac{24}{20} > 1$

This is impossible as $r^2 \leq 1$. So our assumption is wrong.

We now choose equation (*i*) as regression of X on Y and (*ii*) as the regression of Y on X.

Let equation (*i*) be regression equation of X on Y

$$3X + 4Y = 44$$
$$3X = 44 - 4Y$$
$$X = \frac{44}{3} - \frac{4}{3}Y$$

$\therefore$ $bxy = -\frac{4}{3}$ ...(*iii*)

Equation (*ii*) be regression equation of Y on X

$$5X + 8Y = 80$$
$$8Y = 80 - 5X$$
$$Y = \frac{80}{8} - \frac{5}{8}X$$

$\therefore$ $byx = -\frac{5}{8}$ ...(*iv*)

$\therefore$ $r = -\sqrt{bxy \cdot byx} = -\sqrt{\left(\frac{-4}{3}\right) \times \left(\frac{-5}{8}\right)}$

$$= -\sqrt{\frac{5}{6}} = -\sqrt{0.8333} = -0.912 \approx -0.91$$

Calculation of σ_y

We know, $bxy = r \cdot \frac{\sigma_x}{\sigma_y}$

Given: $bxy = \frac{-4}{3}, r = -0.91, \sigma_x^2 = 30 \Rightarrow \sigma_x = \sqrt{30} = 5.47$

Substituting the given values in the formula of bxy, we get

$$-1.33 = -0.91 \times \frac{5.47}{\sigma_y}$$

$$\sigma_y = \frac{0.91 \times 5.47}{1.33} = 3.74$$

Example 35. In a partially destroyed laboratory record of an analysis of correlation data, the following results are legible:

Variance of X = 9

Regression equations

$$8X - 10Y + 66 = 0$$

$$40X - 18Y = 214$$

Find (*i*) $\bar{X}$ and $\bar{Y}$ (*ii*) r_{xy} (*iii*) S.D. of Y.

Solution: **(*i*) Calculation of $\bar{X}$ and $\bar{Y}$**

$$8X - 10Y + 66 = 0$$

$$\Rightarrow \quad 8X - 10Y = -66 \quad \text{...(i)}$$

$$40X - 18Y = 214 \quad \text{...(ii)}$$

Multiplying equation (*i*) by 5 and subtracting (*ii*) from it

$$40X - 50Y = -330$$

$$40X - 18Y = 214$$

$$- \quad + \quad -$$

$$-32Y = -544$$

$$\therefore \quad Y = \frac{544}{32} = 17 \quad \text{or} \quad \bar{Y} = 17$$

By putting the value of Y = 17 in equation (*i*)

$$8X - 10(17) = -66$$

$$8X = -66 + 170$$

$$8X = 104$$

$$\therefore \quad X = 13 \quad \text{or} \quad \bar{X} = 13$$

$$\therefore \quad \bar{X} = 13, \ \bar{Y} = 17$$

(*ii*) Calculation of Coefficient of Correlation (r_{xy})

Let us assume that the first equation be regression equation of Y on X.

$$8X - 10Y + 66 = 0$$

$$-10Y = -66 - 8X$$

or $$10Y = 66 + 8X$$

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

$$Y = 6.6 + 0.8X$$

$$byx = 0.8$$

$\therefore$ Assuming second equation as regression equation of X on Y.

$$40X - 18Y = 214$$

$$40X = 214 + 18Y$$

$$X = \frac{214}{40} + \frac{18}{40} Y$$

$$X = 5.35 + 0.45Y$$

$$bxy = +0.45$$

$$\therefore \quad r = \sqrt{bxy \, . \, byx}$$

Now, $$= \sqrt{0.45 \times 0.8} = +0.6$$

(*iii*) Calculation of σ_y

Given, variance of $X = \sigma_x^2 = 9 \Rightarrow \sigma_x = 3$

$$bxy = 0.45, \ r = +0.6$$

We know, $bxy = r \cdot \frac{\sigma_x}{\sigma_y}$

Substituting the values, we get

$$0.45 = 0.6 \times \frac{3}{\sigma_y}$$

$$\Rightarrow \quad \sigma_y = \frac{1.8}{0.45} = 4$$

Hence, $\bar{X} = 13, \bar{Y} = 17, r = +0.6, \sigma_y = 4.$

Example 36. The two lines of regression are given as follows:

$$Y = -4 + \frac{2}{3} X \qquad X = -5 + \frac{5}{3} Y$$

Find (*i*) Y when $X = 3$ and X when $Y = 3$ (*ii*) r_{xy}

Solution: Given, $$Y = -4 + \frac{2}{3} X \quad \text{...(i)}$$

$$X = -5 + \frac{5}{3} Y \quad \text{...(ii)}$$

Let equation (*i*) as regression of Y on X.

$$Y = -4 + \frac{2}{3} X$$

$\therefore \quad byx = \frac{2}{3}$

Taking equation (*ii*) as regression of X on Y

$$X = -5 + \frac{5}{3}Y$$

$\therefore \quad bxy = \frac{5}{3}$

$\therefore \quad r^2_{xy} = \frac{2}{3} \times \frac{5}{3} = \frac{10}{9} = 1.1 > 1.$

Here, $r^2 > 1$ which is impossible as $r^2 \leq 1$. So our assumption is wrong.

Reversing the assumption and taking equation (*i*) as regression of X on Y

$$Y = -4 + \frac{2}{3}X \implies 3Y = -12 + 2X$$

$$\Rightarrow \quad 2X = 3Y + 12$$

$$\Rightarrow \quad X = \frac{3}{2}Y + 6$$

$\therefore \quad bxy = \frac{3}{2}$ ...(iii)

Taking equation (*ii*) as regression of Y on X

$$X = -5 + \frac{5}{3}Y \quad or \quad 3X = -15 + 5Y$$

$$\Rightarrow \quad 5Y = 3X + 15$$

$$\Rightarrow \quad Y = \frac{3}{5}X + 3$$

$\therefore \quad byx = \frac{3}{5}$ ...(iv)

$\therefore \quad r^2 = bxy \,.\, byx = \frac{3}{2} \times \frac{3}{5} = \frac{9}{10} = 0.9$

Since, $r^2 < 1$, our supposition that equation (*i*) is the line of regression of X on Y and equation (*ii*) is the line of regression of Y on X is true.

(*i*) To obtain an estimate of Y when X = 3, we use the line of regression of Y on X *viz.*, (*ii*) or (*iv*).

Thus, from (*iv*), $Y = \frac{3}{5}X + 3$

Put $X = 3$, $Y = \frac{3}{5}(3) + 3 = \frac{9}{5} + 3 = \frac{24}{5} = 4.8$

To obtain an estimate of X when Y = 3, we use the line of regression of X on Y *viz.* (*i*) or (*iii*).

Thus, from (*iii*), $X = \frac{3}{2}Y + 6$

Put $Y = 3$, $X = \frac{3}{2} \times 3 + 6 = \frac{9}{2} + 6 = \frac{21}{2} = 10.5$

(*ii*) $r_{xy} = \sqrt{bxy \,.\, byx} = \sqrt{\frac{3}{2} \times \frac{3}{5}} = \sqrt{\frac{9}{10}} = \sqrt{0.9} = 0.948$

Example 37. A student obtained the following regression equations. Do you agree with him?

$$6X = 15Y + 21$$
$$21X + 14Y = 56$$

Solution: Here we have two possibilities:

Case I: Treating equation (*i*) as regression equation of X on Y:

$$6X = 15Y + 21$$

or $X = \frac{15}{6}Y + \frac{21}{6}$ ...(i)

Clearly, $bxy = \frac{15}{6}$

Equation (*ii*) as regression equation of Y on X:

$$21X + 14Y = 56$$

or $14Y = 56 - 21X$ ...(ii)

or $Y = \frac{56}{14} - \frac{21}{14}X$

$\therefore \quad byx = -\frac{21}{14}$

Now, $r^2 = bxy.\ byx = \frac{15}{6} \times \frac{-21}{14} < 0$. Here $r^2 < 0$ which is impossible as $r^2 \geq 0$

Case II: Treating equation (*i*) as regression equation of Y on X:

$$6X = 15Y + 21$$

or $15Y = 6X - 21$

or $Y = \frac{6X}{15} - \frac{21}{15}$

$\therefore \quad byx = \frac{6}{15}$

Equation (*ii*) as regression equation of X on Y

$$21X + 14Y = 56$$
$$21X = 56 - 14Y$$
$$X = \frac{56}{21} - \frac{14}{21}Y$$

$$\therefore \quad bxy = -\frac{14}{21}$$

Now, $r^2 = byx \cdot bxy = \frac{6}{15} \times \frac{-14}{21} < 0$. Here $r^2 < 0$ which is impossible as $r^2 \geq 0$.

Here, $r^2 < 0$ which is impossible as $r^2 \geq 0$. Hence, calculations done by the student are wrong.

Example 38. If the regression coefficient of X on Y is $-1/6$ and that of Y on X is $-3/2$. What is the value of correlation coefficient between X and Y?

Solution: Given, $bxy = \frac{-1}{6}$, $byx = \frac{-3}{2}$

$$r = -\sqrt{bxy \cdot byx}$$

$$= -\sqrt{\left(\frac{-1}{6}\right)\left(\frac{-3}{2}\right)} = -\sqrt{\frac{1}{4}} = -\frac{1}{2} = -0.5$$

Hence, $r = -0.5$

IMPORTANT TYPICAL EXAMPLE

Example 39. The regression equation of profits (X) on sales (Y) of a certain firm is $6Y - 10X + 210 = 0$. The average sales of the firm were Rs. 88,000 and the variance of profits is $\frac{16}{25}$th of the variance of sales. Find the average profits and the coefficient of correlation between sales and profits.

Solution: Regression equation of profits (X) on sales (Y) is:

$$6Y - 10X + 210 = 0$$

The average profits can be obtained by putting Y = 88,000 in the regression equation of X on Y as follows:

$$6 \times 88{,}000 - 10X + 210 = 0 \rightarrow 10X = 5{,}28{,}210$$

$\therefore \quad X = 52{,}821 \quad$ or $\quad \bar{X} = 52{,}821$

Also we are given:

Variance of profits $(\sigma_x^2) = \frac{16}{25}$ variance of sales (σ_y^2)

$$\Rightarrow \quad \frac{\sigma_x}{\sigma_y} = \frac{4}{5}$$

$6Y - 10X + 210 = 0$ is the regression of X on Y

$$-10X = -210 - 6Y$$

or $$10X = 210 + 6Y$$

$$X = \frac{210}{10} + \frac{6}{10} Y$$

$$X = 21 + \frac{3}{5} Y.$$

bxy = regression coefficient of X on $Y = \frac{3}{5}$

$\therefore \quad bxy = r \cdot \frac{\sigma_x}{\sigma_y}$

Since

Putting the values, we get

$$\frac{3}{5} = r \times \frac{4}{5}$$

Thus, $\quad r = \frac{3}{4} = 0.75$

Hence, $\bar{X} = 52$, 821, $r = 0.75$

EXERCISE 2.7

1. From the following regression equations:
$$20X - 9Y = 107$$
$$4X - 5Y = -33$$
Calculate $\bar{X}$, $\bar{Y}$ and r. [Ans. $\bar{X} = 13, \bar{Y} = 17, r = 0.6$]

2. Regression equations of two variables X and Y are as follows:
$$2Y - X - 50 = 0$$
$$3Y - 2X - 10 = 0$$
(*i*) Identify which of the two can be called regression of Y on X and X on Y.
(*ii*) Find the means as well as coefficient of correlation between X and Y.
[Ans. (*i*) (1) Y on X and (2) X on Y; (*ii*) $\bar{X} = 130, \bar{Y} = 90, r = 0.866$]

3. The two regression lines are given by:
$$Y = \frac{40}{18} X - \frac{214}{18} \text{ and } X = \frac{10}{8} Y - \frac{66}{8}$$
Find (*i*) Correlation coefficient between X and Y
(*ii*) Y, when $X = 10$
(*iii*) X, when $Y = 10$
(*iv*) σ_y, if $\sigma_x^2 = 9$ [Ans. (*i*) $r = 0.6$, (*ii*) 14.6, (*iii*) 9.85, (*iv*) $\sigma_y = 4$]

4. The lines of regression of a bivariate population are:
$$12X - 15Y + 99 = 0$$
$$64X - 27Y = 373$$
The variance of X is 9.
Find: (*i*) the mean value of X and Y
(*ii*) Correlation coefficient between X and Y, and
(*iii*) the Standard deviation of Y.
[Ans. (*i*) $\bar{X} = 13, \bar{Y} = 17$, (*ii*) $r = +0.58$, (*iii*) $\sigma_y = 4.138$]

5. For certain data, the following regression equations were obtained :
$4X - 5Y + 33 = 0$
$20X - 9Y - 107 = 0$
Estimate Y when X = 20 and X when Y = 20.
[**Hint:** See Example 64] [**Ans.** $Y_{20} = 22.6, X_{20} = 14.35$]

6. Given the regression lines as: $3X + 2Y = 26$ and $6X + Y = 31$, find their point of intersection and interpret it. Also find the correlation coefficient between X and Y.
[**Ans.** Point of intersection of the lines of regression gives the mean values. $(\bar{X} = 4$ and $\bar{Y} = 7)$; $r_{xy} = -0.25$]

7. The two regression lines obtained from certain data were : $Y = X + 5$ and $16X = 9Y - 94$. Find the variance of X, if the variance of Y is 16. Also find the covariance between X and Y.
[**Ans.** $\sigma_x = 3$; Cov $(X, Y) = r\sigma_x\sigma_y = 9$]

8. The line of regression of marks in Statistics (X) on marks in Economics (Y) for a class of 50 boys is $3Y - 5X + 180 = 0$. Average marks in Economics = 44 and variance of marks in Statistics is $\frac{9}{16}$th of variance of marks in Economics. Find (*i*) average marks in statistics (*ii*) Coefficient of correlation between X and Y.
[**Ans.** (*i*) $\bar{X} = 62.4$, (*ii*) $r = 0.8$]

9. Equations of two regression lines in a regression analysis are as follows:
$3X + 2Y = 26$ and $6X + Y = 31$
A student obtained the mean values $\bar{X} = 7, \bar{Y} = 4$ and the value of the correlation coefficient $r = +0.5$. Do you agree with him? If not, suggest your results.
[**Ans.** (*i*) $\bar{X} = 4, \bar{Y} = 7$, (*ii*) $r = -0.5$]

10. For a set of 10 pairs of values of X and Y, the regression line of X on Y is $X - 2Y + 12 = 0$, mean and standard deviation of Y being 8 and 2 respectively. Later it is known that a pair $(X = 3, Y = 8)$ was wrongly recorded and the correct pair detected is $(X = 8, Y = 3)$. Find the correct regression line of X on Y.
[**Hint:** See Example 62] [**Ans.** $X = Y - 1$]

11. A student obtained the two regression equations as
$2X - 5Y - 7 = 0$ and $3X + 2Y - 8 = 0$
Do you agree with him?
[**Ans.** (*i*) $byx = \frac{2}{5}, byx = -\frac{2}{3}$ (*ii*) $byx = \frac{5}{2}, byx = -\frac{3}{2}$
$\therefore$ Equations obtained are wrong]

12. The two lines of regression are given as follows:
$5X - 6Y + 90 = 0$, $15X - 8Y - 130 = 0$
(*i*) Find $\bar{X}$ and $\bar{Y}$ (*ii*) Find r_{xy} (*iii*) Estimate Y when $X = 10$ (*iv*) Estimate X when $Y = 20$.
[**Ans.** $\bar{X} = 30$, $\bar{Y} = 40$, $r = 0.67$, $Y = 23.33$, $X = 19.33$]

STANDARD ERROR OF ESTIMATE

In regression, given the value of independent variable, we estimate the value of dependent variable by using/applying regression equation. To find out an estimate that is 100% accurate is almost impossible. If we want to make sure that to what extent the estimates made by us are accurate or reliable, then this can be done with the help of standard error of estimate. By using standard error of estimate, we can check the reliability of our estimates. Standard error of estimate shows that to what extent the estimated values by regression line are closer to actual values.

For two regression lines (Regression of X on Y and Regression of Y on X), there are two standard error of estimates:

(*i*) **Standard Error of Estimate of Y on X (S_{yx})**

(*ii*) **Standard Error of Estimate of X on Y (S_{xy})**

(*i*) **Standard Error of Estimate of Y on X:** It is denoted by S_{yx}. Its computation is made by the following formulae:

First formula:

$$S_{yx} = \sqrt{\frac{\Sigma (Y - Y_c)^2}{N}}$$

Here, Y = Actual values, Y_c = Estimated values.

Second formula:

$$S_{yx} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}$$

Where, a and b are to be obtained from normal equations and a = intercept, b = slope of line.

Third formula:

$$S_{yx} = \sigma_y \sqrt{1 - r^2}$$

Where, σ_y = SD of Y; r = coefficient of correlation between X and Y.

The third formula is suitable for use when we are given the values of correlation coefficient (r) and standard deviations (σ_x and σ_y).

(*ii*) **Standard Error of Estimate of X on Y:** It is denoted by S_{xy}. Its computation is done by the following formulae:

First formula:

$$S_{xy} = \sqrt{\frac{\Sigma (X - X_c)^2}{N}}$$

Here, X = Actual values, X_c = Estimated values

Second formula:

$$S_{xy} = \sqrt{\frac{\Sigma X^2 - a\Sigma X - b\Sigma XY}{N}}$$

Where, a and b are to be obtained from normal equations and a = intercept, b = slope of line.

Third formula:

$$S_{xy} = \sigma_x \sqrt{1 - r^2}$$

Where, σ_x = SD of Y; r = coefficient of correlation between X and Y.

The third formula is suitable for use when we are given the values of correlation coefficient and standard deviations (σ_x and σ_y).

The following examples make the computation of standard error of estimate more clear:

Example 40. Find the 'standard error of the estimates'.

$\sigma_x = 4.4, \sigma_y = 2.2, r = 0.8$

Solution: Given, $r = 0.8, \sigma_x = 4.4, \sigma_y = 2.2$

As 'r', σ_x and σ_y are known, the following formulae are used to find the 'standard error of estimates':

$$S_{yx} = \sigma_y \sqrt{1 - r^2}$$

$$S_{xy} = \sigma_x \sqrt{1 - r^2}$$

Putting the given values in (*i*) and (*ii*), we get

$$S_{yx} = 2.2 \times \sqrt{1 - (0.8)^2} = 2.2 \times \sqrt{1 - 0.64}$$

$$= 2.2 \times \sqrt{0.36} = 2.2 \times 0.6 = 1.32$$

$$S_{xy} = 4.4 \times \sqrt{1 - (0.8)^2} = 4.4 \times \sqrt{1 - 0.64}$$

$$= 4.4 \times \sqrt{0.36} = 4.4 \times 0.6 = 2.64$$

Example 41. For a set of 10 pairs of reading on X and Y, the coefficient of correlation is 0.856 and the standard deviation of Y is 5.54. Find the standard error of estimate of Y on X.

Solution: We are given:

$r = 0.856$, $\sigma_y = 5.54$

The standard error of estimate (S_{yx}) is given by

$$S_{yx} = \sigma_y \sqrt{1 - r^2}$$

$$= 5.54 \sqrt{1 - (0.856)^2} = 5.54 \sqrt{1 - 0.7327}$$

$$= 5.54 \sqrt{0.2673} = 5.54 \times 0.5170 = 2.864$$

Example 42. From the data given below:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Compute two regression equations and calculate the standard error of the estimate (S_{yx} and S_{xy}).

Calculation of Regression Equations

X	$\bar{X} = 6$ $(X - \bar{X})$ x	x^2	Y	$\bar{Y} = 8$ $(Y - \bar{Y})$ y	y^2	xy
6	0	0	9	1	1	0
2	−4	16	11	3	9	−12
10	4	16	5	−3	9	−12
4	−2	4	8	0	0	0
8	2	4	7	−1	1	−2
$N = 5$ $\Sigma X = 30$	$\Sigma x = 0$	$\Sigma x^2 = 40$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 20$	$\Sigma xy = -26$

We have, $\bar{X} = \frac{30}{5} = 6 \quad \bar{Y} = \frac{40}{5} = 8$

$$bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{-26}{20} = -1.3$$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{-26}{40} = -0.65$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 8 = -0.65(X - 6)$$

$$Y - 8 = -0.65X + 3.9$$

$$Y = -0.65X + 11.9$$

$$Y = 11.9 - 0.65X$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

$$X - 6 = -1.3(Y - 8)$$

$$X - 6 = -1.3Y + 10.4$$

$$X = -1.3Y + 10.4 + 6$$

$$X = -1.3Y + 16.4$$

$$X = 16.4 - 1.3Y$$

Thus, the two regression equations are:

$Y = 11.9 - 0.65X$

$X = 16.4 - 1.3Y$

Calculation of Standard Error of Estimates

From the regression equation of Y on X ($Y_c = 11.9 - 0.65X$) for various values of X, we can find out the corresponding value of Y_c values and from the equation of X on Y ($X_c = 16.4 - 1.3Y$), we can find X_c. These values are:

Computation of Standard Error of Estimate

X	Y	Y_c	X_c	$(Y-Y_c)^2$	$(X-X_c)^2$
6	9	8.0	4.7	1.00	1.69
2	11	10.6	2.1	0.16	0.01
10	5	5.4	9.9	0.16	0.01
4	8	9.3	6.0	1.69	4.00
8	7	6.7	7.3	0.09	0.49
				$\Sigma (Y-Y_c)^2 = 3.10$	$\Sigma(X-X_c)^2 = 6.20$

Standard Error of Estimates:

$$S_{yx} = \sqrt{\frac{\Sigma (Y-Y_c)^2}{N}} = \sqrt{\frac{3.10}{5}} = +0.7874$$

$$S_{xy} = \sqrt{\frac{\Sigma (X-X_c)^2}{N}} = \sqrt{\frac{6.20}{5}} = +1.11$$

Aliter: S_{yx} and S_{xy} can also be calculated as:

$$\sigma_y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{N}} = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{20}{5}} = \sqrt{4} = 2$$

$$\sigma_x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{40}{5}} = \sqrt{8} = 2.828$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{-26}{\sqrt{20\times 40}} = \frac{-26}{\sqrt{800}} = \frac{-26}{28.28} = -0.919$$

$$S_{yx} = \sigma_y\sqrt{1-r^2} = 2\sqrt{1-(-0.919)^2} = 2\times\sqrt{0.155} = 0.7874$$

$$S_{xy} = \sigma_x\sqrt{1-r^2} = 2.828\sqrt{1-(0.919)^2} = 2.828\times\sqrt{0.155} = 1.11$$

Example 43. Given that

$\Sigma Y = 15$, $\Sigma Y = 110$, $\Sigma XY = 400$, $\Sigma X^2 = 250$, $\Sigma Y^2 = 3200$, $N = 10$

(*i*) Compute the regression equation of Y on X

(*ii*) Standard Error of Estimate S_{yx}.

Solution: (*i*) byx or $b = \dfrac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N.\Sigma X^2 - (\Sigma X)^2}$

$$= \frac{10(400)-(15)(110)}{10(250)-(15)^2} = 1.033$$

$$\bar{X} = \frac{15}{10} = 1.5, \quad \bar{Y} = \frac{110}{10} = 11$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 11 = 1.033(x - 1.5)$$

$$Y - 11 = 1.033X - 1.5495$$

$$Y = 9.4505 + 1.033X$$

(*ii*) Standard error of estimate (S_{yx}) is given by

$$S_{yx} = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{N}}$$

$$= \sqrt{\frac{3200-(9.45)(110)-(1.033)(400)}{10}}$$

$$= \sqrt{\frac{3200-1039.5-413.2}{10}} = \sqrt{\frac{1747.3}{10}}$$

$$= 13.21$$

Example 44. From the following data, find the standard error of the estimate of X on Y and Y on X:

X:	1	2	3	4	5
Y:	6	8	7	6	8

Solution:

X	$(X-\bar{X})$ x	x^2	Y	$(Y-\bar{Y})$ y	y^2	xy
1	−2	4	6	−1	1	2
2	−1	1	8	+1	1	−1
3	0	0	7	0	0	0
4	1	1	6	−1	1	−1
5	2	4	8	+1	1	2
$\Sigma X = 15$		$\Sigma x^2 = 10$	$\Sigma Y = 35$		$\Sigma y^2 = 4$	$\Sigma xy = 2$

$$\bar{X} = \frac{15}{5} = 3 \qquad \bar{Y} = \frac{35}{5} = 7$$

$$\sigma_x = \sqrt{\frac{\Sigma x^2}{N}} = \sqrt{\frac{10}{5}} = 1.414$$

$$\sigma_y = \sqrt{\frac{\Sigma y^2}{N}} = \sqrt{\frac{4}{5}} = 0.894$$

$$r = \frac{\Sigma xy}{N.\sigma_x.\sigma_y}$$

$$r = \frac{2}{5\times 1.414\times 0.894} = \frac{2}{6.321} = 0.316$$

$$S_{xy} = \sigma_x \sqrt{1-r^2}$$
$$= 1.414\sqrt{1-(0.316)^2} = 1.414 \times 0.949 = 1.342$$
$$S_{yx} = \sigma_y \sqrt{1-r^2}$$
$$= 0.894\sqrt{1-(0.316)^2} = 0.894 \times 0.949 = 0.848$$

Example 45. For 10 observations on price (X) and supply (Y) the following data were obtained:
$\Sigma X = 130,\ \Sigma Y = 220,\ \Sigma X^2 = 2288$
$\Sigma Y^2 = 5506,\ \Sigma XY = 3467,\ N = 10$
Obtain the standard error of estimate of X on Y and Y on X.

Solution: Given, $N = 10,\ \Sigma X = 130,\ \Sigma Y = 220,\ \Sigma X^2 = 2288$
$\Sigma Y^2 = 5506,\ \Sigma XY = 3467$

$$\sigma_x = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} \quad \text{(Formula of S.D.)}$$
$$= \sqrt{\frac{2288}{10} - \left(\frac{130}{10}\right)^2} = \sqrt{228.8 - 169} = \sqrt{59.8} = 7.73$$
$$\sigma_y = \sqrt{\frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2}$$
$$= \sqrt{\frac{5506}{10} - \left(\frac{220}{10}\right)^2} = \sqrt{550.6 - 484} = \sqrt{66.6} = 8.16$$
$$r = \frac{N.\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{N.\Sigma X^2 - (\Sigma X)^2}\sqrt{N.\Sigma Y^2 - (\Sigma Y)^2}}$$
$$= \frac{10 \times 3467 - (130)(220)}{\sqrt{10 \times 2288 - (130)^2}\sqrt{10 \times 5506 - (220)^2}}$$
$$= \frac{34670 - 28600}{\sqrt{22880 - 16900}\sqrt{55060 - 48400}}$$
$$= \frac{6070}{\sqrt{5980}\sqrt{6660}} = \frac{6070}{6310.84} = 0.961$$

Standard Error of Y on X

$$S_{yx} = \sigma_y\sqrt{1-r^2} = 8.16\sqrt{1-(0.961)^2}$$
$$= 8.16\sqrt{0.07647} = 2.256$$

Standard Error of X on Y

$$S_{xy} = \sigma_x\sqrt{1-r^2}$$
$$= 7.73\sqrt{1-(0.961)^2}$$
$$= 7.73\sqrt{0.07647}$$
$$= 2.137$$

EXERCISE 2.8

1. Find the standard error of estimates:
$\sigma_x = 1.414,\ \sigma_y = 0.894,\ r = 0.316$
[Ans. S_{yx}=.848, S_{xy}=1.342]

2. For a set of 8 pair reading on X and Y, the coefficient of correlation is 0.65 and the standard deviation of Y series is 4.2. Find the standard error of Y on X.
[Ans. S_{yx}=3.1917]

3. From the following data, compute standard error of estimate of the regression of Y on X:
$\Sigma x^2 = 10,\ \Sigma y^2 = 4, \Sigma xy = 2,\ N = 5,$ where $x = X - \overline{X},\ \ y = Y - \overline{Y}$
[Ans. S_{yx}=0.848]

4. Given the following data:

X:	1	2	3	4	5
Y:	2	4	5	3	6

Obtain the two regression equations and calculate the standard error of estimates.
[Ans. X = 0.74Y + 0.2, Y= 0.7 +1.9, S_{xy} = 1.01, S_{yx} = 1.01]

5. Family income and its percentage spent on food in the case of hundred families gave the following bivariate frequency distribution.

Food Expenditure (in %)	Family Income (Rs.)				
	200—300	300—400	400–500	500—600	600—700
10—15	—	—	—	3	7
15—20	—	4	9	4	3
20—25	7	6	12	5	—
25—30	3	10	19	8	—

Obtain the equations of the two lines of regression. Also compute the standard error of the estimates.
[Hint : See Example 55]
[Ans. $Y = -0.02X + 31.5$; $X = -9.6Y + 666$; $S_{yx} = 4.494$, $S_{xy} = 98.47$]

Explained and Unexplained Variation

The total variation in the dependent variable Y can be split into two:

(a) Explained Variation: The variation in Y which is explained by the variation in X is called explained variation in Y.

(b) Unexplained Variation: The variation in Y which is unexplained by the variation in variable X and is due to some other factors (variables) is called unexplained variation in Y.

Symbolically,

Total variation in Y = Explained variation in Y + Unexplained variation in Y

$$\Sigma(Y-\bar{Y})^2 = \Sigma(Y_c-\bar{Y})^2 + \Sigma(Y-Y_c)^2$$

Where, Y_c = computed (or estimated) value of Y on the basis of regression equation

$\bar{Y}$ = Mean value of Y series

Y = Original value of Y series.

A similar relationship we may have for X variable (Dependent) in terms of Y:

$$\Sigma(X-\bar{X})^2 = \Sigma(X_c-\bar{X})^2 + \Sigma(X-X_c)^2$$

Coefficient of Determination: Based on the above expression, the coefficient of determination (r^2) is defined as the ratio of the explained variation to total variation, *i.e.*,

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\Sigma(Y_c-\bar{Y})^2}{\Sigma(Y-\bar{Y})^2}$$

It is clear that the object of coefficient of determination is to determine the percentage variation in Y which is explained by variation in X. For example, let us suppose that the correlation coefficient between X and Y is +0.8, then coefficient of determination $(r^2) = (0.8)^2 = 0.64$. It means that 64% variations in Y are due to variation in X and 36% variations are due to other factors. Thus, explained variations are 64% and unexplained variations are 36%.

Coefficient of Non-Determination: The proportion of unexplained variation to total variation is termed as coefficient of non-determination. It is denoted by k^2, where $k^2 = 1-r^2$. It is also written as:

$$k^2 = \frac{\text{Unxplained variation}}{\text{Total variation}} = 1-r^2$$

The square root of k^2 is termed as coefficient of alientation, *i.e.*, $k = \sqrt{k^2} = \sqrt{1-r^2}$

Standard Error of Estimate: Standard error of estimates of Y on X and X on Y can also be calculated as:

$$S_{yx} = \sqrt{\frac{\Sigma(Y-Y_c)^2}{N}} = \sqrt{\frac{\text{Unexplained variation in } Y}{N}}$$

$$S_{xy} = \sqrt{\frac{\Sigma(X-X_c)^2}{N}} = \sqrt{\frac{\text{Unexplained variation in } X}{N}}$$

Example 46. Given: Explained variation = 19.22
Unexplained variation = 19.70
Determine the coefficient of correlation.

Solution: Total variation = Explained variation + Unexplained variation
= 19.22 + 19.70 = 38.92

$$\text{Coefficient of Determination } (r^2) = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{19.22}{38.92} = 0.4938$$

$$\Rightarrow \text{Coefficient of Correlation } (r) = \sqrt{0.4938} = 0.70$$

Example 47. In fitting of a regression of Y on X to a bivariate distribution consisting of 9 observations, the explained and unexplained variations were computed as 24 and 36 respectively. Find: (*i*) coefficient of determination, and (*ii*) standard error of estimate of Y on X.

Solution: Total variation in Y = Explained variation in Y + Unexplained variation in Y
= 24 + 36 = 60

$$\text{Coefficient of Determination } (r^2) = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{24}{60} = \frac{4}{10} = 0.40$$

$$\text{Standard Error of Estimate of Y on X } (S_{yx}) = \sqrt{\frac{\text{Unexplained variation}}{N}}$$

$$= \sqrt{\frac{36}{9}} = \sqrt{4} = 2$$

Example 48. Given the following data:

X :	1	2	3	4	5
Y :	10	20	30	50	40

Calculate:
(*i*) Regression equation of Y on X.
(*ii*) Total variation in Y.
(*iii*) Unexplained variation in Y.
(*iv*) Explained variation in Y.
(*v*) Standard error of estimate.
(*vi*) Coefficient of determination.

Solution:

X	$x = X-\bar{X}$	x^2	Y	$y = Y-\bar{Y}$	y^2	xy
1	−2	4	10	−20	400	40
2	−1	1	20	−10	100	10
3	0	0	30	0	0	0
4	1	1	50	20	400	20
5	2	4	40	10	100	20
$\Sigma X = 15$ $N = 5$	$\Sigma x = 0$	$\Sigma x^2 = 10$	$\Sigma Y = 150$	$\Sigma y = 0$	$\Sigma y^2 = 1000$	$\Sigma xy = 90$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{15}{5} = 3 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{150}{5} = 30$$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{90}{10} = 9$$

(i) Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 30 = 9\,(X - 3)$$
$$Y - 30 = 9X - 27$$
$$Y = 9X + 3$$

(ii) Total variation in $Y = \Sigma(Y - \bar{Y})^2 = \Sigma y^2 = 1000$

(iii) Unexplained variation in Y:

X	Y	$Y_c = 9X + 3$	$Y - Y_c$	$(Y - Y_c)^2$
1	10	12	−2	4
2	20	21	−1	1
3	30	30	0	0
4	50	39	11	121
5	40	48	−8	64
				$\Sigma(Y - Y_c)^2 = 190$

Unexplained variation $= \Sigma(Y - Y_c)^2 = 190$

(iv) Explained variation in Y = Total variation − Unexplained variation
$= 1000 - 190 = 810$

(v) Standard error of estimate $(S_{yx}) = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}} = \sqrt{\frac{190}{5}} = 6.164$

(vi) Coefficient of determination $(r^2) = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{810}{1000} = 0.81$

Example 49. From the following data:

Age of husband (years):	18	19	20	21	22	23	24	25	26	27
Age of wife (years):	17	17	18	18	18	19	19	20	21	21

Obtain the following:

(i) The regression of age of husband on the age of wife.

(ii) Total variation in the age of husband.

(iii) The magnitude of variation in age of the husband, explained by the regression equation.

(iv) Standard error of the estimate of age of husband.

Let Y denote age of husband and X denote age of wife.

Y	A = 23 dy	dy^2	X	A = 19 dx	dx^2	$dxdy$
18	−5	25	17	−2	4	10
19	−4	16	17	−2	4	8
20	−3	9	18	−1	1	3
21	−2	4	18	−1	1	2
22	−1	1	18	−1	1	1
23 = A	0	0	19 = A	0	0	0
24	+1	1	19	0	0	0
25	+2	4	20	+1	1	2
26	+3	9	21	+2	4	6
27	+4	16	21	+2	4	8
$\Sigma Y = 225$ N = 10 $\therefore \bar{Y} = 22.5$	$\Sigma dy = -5$	$\Sigma dy^2 = 85$	$\Sigma X = 188$ N = 10 $\therefore \bar{X} = 18.8$	$\Sigma dx = -2$	$\Sigma dx^2 = 20$	$\Sigma dxdy = 40$

(i) $$byx = \frac{N.\Sigma dxdy - \Sigma dx.\Sigma dy}{N.\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{10 \times 40 - (-2)(-5)}{10 \times 20 - (-2)^2} = \frac{400 - 10}{200 - 4} = \frac{390}{196} = 1.989 \cong 1.99$$

$$\sigma_y^2 = \frac{\Sigma dy^2}{N} - \left(\frac{\Sigma dy}{N}\right)^2$$

$$= \frac{85}{10} - \left(\frac{-5}{10}\right)^2$$

$$= 8.5 - 0.25 = 8.25$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 22.5 = 1.99(X - 18.8)$$
$$Y - 22.5 = 1.99X - 37.412$$
$$\therefore \quad Y = 1.99X - 14.912$$

(ii) Total variation in $Y = \Sigma(Y - \bar{Y})^2 = N.\sigma_y^2$

$$= 10 \times 8.25 = 82.5$$

(iii) **Explained variation in Y :**

X	Y	$Y_e = (1.99X - 14.912)$	$Y_e - \bar{Y}$	$(Y_e - \bar{Y})^2$
17	18	18.9	−3.6	12.96
17	19	18.9	−3.6	12.96
18	20	20.9	−1.6	2.56
18	21	20.9	−1.6	2.56
18	22	20.9	−1.6	2.56
19	23	22.9	0.4	0.16
19	24	22.9	0.4	0.16
20	25	24.9	2.4	5.76
21	26	26.9	4.4	19.36
21	27	26.9	4.4	19.36
				$\Sigma(Y_e - \bar{Y})^2 = 78.40$

Explained variation in Y = 78.40

∴ Magnitude of variation in age of husband (Y) explained by the regression equation = 78.40

Unexplained variation in Y = Total variation − Explained variation
= 82.5 − 78.40 = 4.1

(iv) **Standard Error of Estimate (S_{yx})**

$$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}} = \sqrt{\frac{\text{Unexplained Variation}}{N}} = \sqrt{\frac{4.1}{10}} = 0.64$$

EXERCISE 2.9

1. The coefficient of correlation (r) between two variables X and Y is + 0.95. What percent variation in Y (dependent variable) remains unexplained by the variation in X (the independent variable).
[Hint: $r^2 = 0.9025$, Explained variation = 90.25%]
[**Ans.** Unexplained variation = 9.75%]

2. If the explained variation is 15.24 and the unexplained variation is 27.09, find the coefficient of determination.
[Hint: r^2 = Explained Variation ÷ Total Variation] [Ans. r^2=0.36]

3. Given the bivariate data:

X:	1	5	3	2	1	1	7	3
Y:	6	1	0	0	1	2	1	5

Obtain: *(i)* Regression equation of Y on X
(ii) Total variation in Y, and

(iii) Explained variation in Y
(iv) Standard error of estimate of Y on X.
[**Ans.** *(i)* Y = 2.874 − 0.304X *(ii)* 38 *(iii)* 3.042 (iv) 2.0299]

4. The coefficient of correlation (r) between consumption expenditure (C) and disposable income (I) in a study was found to be +0.8. What percentage of variation in C are explained by variations in Y? [**Ans.** 64%]

5. From the following data, find out (i) correlation coefficient (ii) linear regression line of Y on X. Also find the percentage of variation explained by the regression line.

X:	1	2	3	4	5
Y:	2	5	3	8	7

[Hint: See Example 59]
[**Ans.** $r = 0.806$, $Y = 1.3X + 1.1$,
$r^2 = 0.6496 \Rightarrow 64.96\%$ variations are explained by the regression line.]

6. Given the following data:
$S_{yx} = 35$, $\Sigma Y^2 = 800$, $\bar{Y} = 5$, $N = 20$
Find (i) total variation in Y which are unexplained and explained by X (ii) coefficient of determination and (iii) coefficient of non-determination.
[**Ans.** *(i)* 300, 245, 55, *(ii)* 0.1833, *(iii)* 0.8167]

MISCELLANEOUS SOLVED EXAMPLES

Example 50. From the following data, obtain the two regression equations:

X:	1	2	3	4	5	6	7	8	9
Y:	9	8	10	12	11	13	14	16	15

Verify that the coefficient of correlation is the geometric mean of the two regression coefficients.

Solution: *(i)* **Calculation of Regression Equations**

X	$\bar{X}=5$ x	x^2	Y	$\bar{Y}=12$ y	y^2	xy
1	−4	16	9	−3	9	+12
2	−3	9	8	−4	16	+12
3	−2	4	10	−2	4	+4
4	−1	1	12	0	0	0
5	0	0	11	−1	1	0
6	+1	1	13	+1	1	+1
7	+2	4	14	+2	4	+4
8	+3	9	16	+4	16	+12
9	+4	16	15	+3	9	+12
$\Sigma X = 45$	$\Sigma x = 0$	$\Sigma x^2 = 60$	$\Sigma Y = 108$	$\Sigma y = 0$	$\Sigma y^2 = 60$	$\Sigma xy = 57$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{45}{9} = 5 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{108}{9} = 12$$

Since the actual means of X and Y are whole numbers, therefore we should take the deviations from $\bar{X}$ and $\bar{Y}$.

$$bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{57}{60} = +0.95$$

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{57}{60} = +0.95$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 12 = +0.95\,(X - 5)$$
$$Y - 12 = 0.95X - 4.75$$
$$Y = 0.95X + 7.25$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 5 = +0.95\,(Y - 12)$$
$$X - 5 = 0.95Y - 11.4$$
$$X = 0.95Y - 6.4$$

Calculation of Coefficient of Correlation

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$
$$= \frac{57}{\sqrt{60}\sqrt{60}} = +0.95$$

Verification:

$$r = \sqrt{bxy \cdot byx}$$
$$= \sqrt{0.95 \times 0.95}$$
$$= 0.95$$

Hence the result is verified.

Example 51. The following table gives the ages and blood pressure of 10 women:

Age:	56	42	36	47	49	42	60	72	63	55
Blood Pressure:	147	125	118	128	145	140	155	160	149	150

Estimate the blood pressure of a woman whose age is 45 years.

Let age be denoted by X and blood pressure be denoted by Y.

X	$A = 49$, dx	dx^2	Y	$A = 140$, dy	dy^2	$dxdy$
56	+7	49	147	+7	49	+49
42	−7	49	125	−15	225	+105
36	−13	169	118	−22	484	+286
47	−2	4	128	−12	144	+24
49 = A	0	0	145	+5	25	0
42	−7	49	140 = A	0	0	0
60	+11	121	155	+15	225	+165
72	+23	529	160	+20	400	+460
63	+14	196	149	+9	81	+126
55	+6	36	150	+10	100	60
$\Sigma X = 522$	$\Sigma dx = 32$	$\Sigma dx^2 = 1202$	$\Sigma Y = 1417$	$\Sigma dy = 17$	$\Sigma dy^2 = 1733$	$\Sigma dxdy = 1275$

For estimating the blood pressure of a woman whose age is 45, we fit a regression equation of Y on X:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{522}{10} = 52.2 \qquad \bar{Y} = \frac{\Sigma N}{N} = \frac{1417}{10} = 141.7$$

$$byx = \frac{N.\Sigma dxdy - \Sigma dx\,.\,\Sigma dy}{N.\Sigma dx^2 - (\Sigma dx)^2}$$
$$= \frac{10 \times 1275 - (32)(17)}{10 \times 1202 - (32)^2} = \frac{12750 - 544}{12020 - 1024} = \frac{12206}{10996} = 1.11$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 141.7 = 1.11(X - 52.2)$$
$$Y - 141.7 = 1.11X - 57.942$$
$$Y = 1.11X + 83.758$$

For $X = 45$, $Y = 1.11\,(45) + 83.758$

$$= 49.95 + 83.758 = 133.708$$
$$= 133.708 \approx 134$$

Thus, the blood pressure of a woman whose age is 45 = 134.

Example 52. On each of 30 items, two measurements on X and Y are made. The following summations are given:

$$\Sigma X = 15,\ \Sigma Y = -6,\ \Sigma XY = 56,\ \Sigma X^2 = 61 \text{ and } \Sigma Y^2 = 90.$$

Calculate the product moment correlation coefficient and the slope of the regression line of Y on X. How would your results be affected if X is replaced by $U = \frac{X - 1}{2}$?

Solution: (*i*) **Coefficient of Correlation**

$$r = \frac{30\times 56+15\times 6}{\sqrt{30\times 61-(15)^2}\sqrt{30\times 90-(-6)^2}} = \frac{1770}{\sqrt{1605}\times\sqrt{2664}} = 0.856$$

Since r is independent of change of scale and of change of origin, it would remain same even after making the above mentioned transformation.

(*ii*) **Regression Coefficient of Y on X**

$$byx = \frac{1770}{1605} = 1.10$$

We know that byx is independent of change of origin but not of change of scale. If byx denote regression coefficient of Y on X and bvu denote regression coefficient of v on u where, $u = \frac{X-A}{h}$ and $v = \frac{Y-B}{k}$, we know that $byx = \frac{k}{h} bvu$

or $bvu = \frac{h}{k} byx$.

In the example, it is given that $h = 2$ and $k = 1$, *i.e.*, $bvu = 2byx$.

Hence, the new regression coefficient will be two times the old regression coefficient of, *i.e.*, regression coefficient of Y on $U(=\frac{X-1}{2})$ will be equal to 2×1.10, *i.e.*, 2.20.

Example 53. For certain data, $3X + 2Y - 26 = 0$ and $6X + Y - 31 = 0$ are the two regression equations. Find the values of means and coefficient of correlation.

Solution: **Calculation of $\bar{X}$ and $\bar{Y}$**

$$3X + 2Y - 26 = 0 \quad ...(i)$$
$$6X + Y - 31 = 0 \quad ...(ii)$$

Multiplying (*ii*) by 2 and subtracting (*i*) from (*ii*)

$$12X + 2Y - 62 = 0$$
$$3X + 2Y - 26 = 0$$
$$- \quad - \quad +$$
$$9X - 36 = 0$$
$$9X = 36$$

$\therefore X = 4$ or $\bar{X} = 4$

Putting the value of $X = 4$ in (*i*)

$$3(4) + 2Y - 26 = 0$$
$$12 + 2Y - 26 = 0$$
$$2Y = 14$$
$$Y = 7 \text{ or } \bar{Y} = 7$$

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

Calculation of Coefficient of Correlation

Let us take equation (*i*) as Y on X and (*ii*) as X on Y

Regression Equation of Y on X

$$3X + 2Y - 26 = 0$$
$$2Y = 26 - 3X$$
$$Y = \frac{26}{2} - \frac{3}{2}X$$
$$byx = -\frac{3}{2} \quad ...(i)$$

$\therefore$

Regression Equation of X on Y

$$6X + Y - 31 = 0$$
$$6X = 31 - Y$$
$$X = \frac{31}{6} - \frac{1}{6}Y$$
$$bxy = -\frac{1}{6} \quad ...(ii)$$

$\therefore$

Now,
$$r = \sqrt{byx \cdot bxy}$$
$$= -\sqrt{\left(\frac{-3}{2}\right)\left(\frac{-1}{6}\right)} = -\sqrt{\left(\frac{3}{12}\right)} = -\sqrt{\left(\frac{1}{4}\right)} = \frac{-1}{2} = -0.50$$

Example 54. The following results were worked out from the scores in Statistics and Mathematics in a certain examination:

	Score in Statistics (X)	Score in Mathematics (Y)
Mean:	39.5	47.5
Standard Deviation:	10.8	17.8
Coefficient of correlation = +0.42		

Find both the regression equations. Use these regressions to estimate the value of Y for X=50 and also estimate the value of X for Y=30.

Solution: Given, $\bar{X} = 39.5$, $\bar{Y} = 47.5$, $\sigma_x = 10.8$, $\sigma_y = 17.8$, $r = +0.42$

(*i*) **Regression Equation of X on Y**

$$X - \bar{X} = r \cdot \frac{\sigma_x}{\sigma_y}(Y - \bar{Y})$$
$$X - 39.5 = 0.42 \times \frac{10.8}{17.8}(Y - 47.5)$$
$$X - 39.5 = 0.25(Y - 47.5) \Rightarrow X = 0.25Y + 27.625$$

Regression Equation of Y on X

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

$$Y - 47.5 = +0.42 \times \frac{17.8}{10.8}(X - 39.5)$$

$$Y - 47.5 = 0.69\,(X - 39.5) \quad \Rightarrow \quad Y = 0.69X + 20.245$$

(*ii*) For X = 50, Y = 0.69 (50) + 20.245 = 54.745

For Y = 30, X = 0.25 (30) + 27.625 = 35.125

Example 55. Family income and its percentage spent on food in case of hundred families gave the following bivariate distribution:

Food Expenditure (in %)	Family Income (Rs)				
	200—300	300—400	400—500	500—600	600—700
10—15	—		—	3	7
15—20	—	4	9	4	3
20—25	7	6	12	5	—
25—30	3	10	19	8	—

Obtain the equations of two lines of regression. Also compute standard error of estimates.

Solution: **(Landscape Table Given at Page 157)**

Let $dx = \frac{X - 450}{100}$ and $dy = \frac{Y - 17.5}{5}$

Regression Coefficient of Y and X

$$byx = \frac{N \cdot \Sigma fdxdy - \Sigma fdx\, \Sigma fdy}{N \cdot \Sigma fdx^2 - (\Sigma fdx)^2} \times \frac{i_y}{i_x}$$

$$= \frac{100(-48) - 0 \times 100}{100 \times 120 - (0)^2} \times \frac{5}{100} = \frac{-4800}{12000} \times \frac{1}{20} = \frac{-2}{100} = -0.02$$

Regression Coefficient of X on Y

$$bxy = \frac{N \cdot \Sigma fdxdy - \Sigma fdx\, \Sigma fdy}{N \cdot \Sigma fdy^2 - (\Sigma fdy)^2} \times \frac{i_x}{i_y}$$

$$= \frac{100(-48) - 0 \times 100}{100 \times 200 - (100)^2} \times \frac{100}{5} = \frac{-4800 - 0}{20000 - 10000} \times 20$$

$$= \frac{-4800}{10000} \times 20 = \frac{-48}{5} = -9.6$$

$$\bar{X} = A + \frac{\Sigma fdx}{N} \times i_x = 450 + \frac{0}{100} \times 100 = 450$$

$$\bar{Y} = A + \frac{\Sigma fdy}{N} \times i_y = 17.5 + \frac{100}{100} \times 5 = 22.5$$

Solution 55.

Y ↓	M.V.		dy \ dx	200—300	300—400	400—500	500—600	600—700	f	fdy	fdy²	fdxdy
			M.V.	250	350	450	550	650				
				−200	− 100	0	100	200				
			dx	− 2	− 1	0	1	2				
10—15	12.5	−5	−1	—	—	—	−1, 3, −3	−2, 7, −14	10	−10	10	− 17
15—20	17.5	0	0	—	0, 4, 0	0, 9, 0	0, 4, 0	0, 3, 0	20	0	0	0
20—25	22.5	+5	+1	−2, 7, −14	−1, 6, −6	0, 12, 0	1, 5, 5	—	30	30	30	− 15
25—30	27.5	10	+2	−4, 3, −12	−2, 10, −20	0, 19, 0	2, 8, 16	—	40	80	160	− 16
			f	10	20	40	20	10	N = 100	100	200	− 48
			fdx	−20	− 20	0	20	20	0			
			fdx²	40	20	0	20	40	120			
			fdxdy	− 26	− 26	0	18	− 14	− 48			

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 22.5 = -0.02(X - 450)$$
$$Y = -0.02X + 9 + 22.5$$
$$Y = -0.02X + 31.5$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 450 = -9.6\,(Y - 22.5)$$
$$X - 450 = -9.6Y + 216$$
$$X = -9.6Y + 666$$

Standard Error of Estimates

$$\sigma_x = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2} \times i_x$$
$$= \sqrt{\frac{120}{100} - \left(\frac{0}{100}\right)^2} \times 100$$
$$= 109.545$$

$$\sigma_y = \sqrt{\frac{\Sigma fdy^2}{N} - \left(\frac{\Sigma fdy}{N}\right)^2} \times i_y$$
$$= \sqrt{\frac{200}{100} - \left(\frac{100}{100}\right)^2} \times 5$$
$$= \sqrt{2-1} \times 5 = 5$$

$$r^2 = byx \,.\, bxy = (-0.02) \times (-9.6) = 0.192$$
$$S_{yx} = \sigma_y \sqrt{1-r^2} = 5 \times \sqrt{1-0.192} = 5 \times \sqrt{0.808} = 4.494$$
$$S_{xy} = \sigma_x \sqrt{1-r^2} = 109.545 \times \sqrt{1-0.192} = 109.545 \times \sqrt{0.808} = 98.47$$

Example 56. Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations:

$$2Y - X - 50 = 0$$
$$3Y - 2X - 10 = 0$$

Also calculate the standard error of estimate of Y on X, given that the standard deviation of X is 3.

Solution: **(a) Calculation of $\bar{X}$ and $\bar{Y}$**

$$2Y - X - 50 = 0 \quad \text{...(i)}$$
$$3Y - 2X - 10 = 0 \quad \text{...(ii)}$$

Multiplying (*i*) by 2 and subtracting (*ii*) from it,

$$4Y - 2X - 100 = 0$$
$$3Y - 2X - 10 = 0$$
$$- \quad + \quad +$$
$$Y - 90 = 0$$

or $Y = 90$ or $\bar{Y} = 90$

Putting the value of Y in (*i*)

$$2(90) - X - 50 = 0$$
$$180 - X - 50 = 0$$
$$\therefore X = 130 \text{ or } \bar{X} = 130$$
$$\therefore \bar{X} = 130,\ \bar{Y} = 90$$

(b) Calculation of Correlation Coefficient

Let us assume equation (*i*) as Y on X and equation (*ii*) as X on Y.

Regression of Y on X

$$2Y - X - 50 = 0$$
$$2Y = 50 + X$$
$$Y = 25 + \frac{1}{2}X$$

or

$$\therefore \quad byx = \frac{1}{2}$$

Regression of X on Y

$$3Y - 2X - 10 = 0$$
$$3Y - 10 = 2X$$

or

$$2X = 3Y - 10$$
$$X = -5 + \frac{3}{2}Y$$
$$\therefore \quad bxy = \frac{3}{2}$$

$$r^2 = byx \,.\, bxy = \frac{1}{2} \times \frac{3}{2}$$
$$\therefore \quad r = \sqrt{\frac{3}{4}} = \sqrt{0.75} = +0.866$$

(*c*) We know that $byx = r.\frac{\sigma_y}{\sigma_x}$

$$\Rightarrow \quad \frac{1}{2} = 0.866.\frac{\sigma_y}{3} \qquad [\because \sigma_x = 3 \text{ given}]$$
$$\Rightarrow \quad \sigma_y = 1.732$$

Standard error of estimate of Y on X is

$$S_{yx} = \sigma_y \sqrt{1-r^2}$$

or

$$S_{yx} = 1.732 \sqrt{1 - \frac{3}{4}} = 1.732 \times \frac{1}{2} = 0.866$$

Example 57. A departmental store gives in-service training to its salesmen which is followed by a test. It is considering whether it should terminate the service of any salesmen who does not do well in the test. The following data give the test scores and sales made by nine salesmen during a certain period :

Test Scores	14	19	24	21	26	22	15	20	19
Sales ('00 Rs.)	31	36	48	37	50	45	33	41	39

Calculate the coefficient of correlation between the test scores and the sales. Does it indicate that the termination of services of low test scores is justified ? If the firm wants a minimum sales volume of Rs. 3,000, what is the minimum test score that will ensure continuation of service ? Also estimate the most probable sales volume of a salesman making a score of 28.

Solution: Let X denote the test scores of the salesmen and Y denote their corresponding sales (in '00 Rs.)

Calculations for Regression Lines

X	Y	$x = X - \bar{X} = X - 20$	$y = Y - \bar{Y} = Y - 40$	x^2	y^2	xy
14	31	– 6	– 9	36	81	54
19	36	– 1	– 4	1	16	4
24	48	4	8	16	64	32
21	37	1	– 3	1	9	– 3
26	50	6	10	36	100	60
22	45	2	5	4	25	10
15	33	– 5	– 7	25	49	35
20	41	0	1	0	1	0
19	39	– 1	– 1	1	1	1
$\Sigma X = 180$	$\Sigma Y = 360$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 120$	$\Sigma y^2 = 346$	$\Sigma xy = 193$

$$\bar{X} = \frac{\Sigma X}{N} = \frac{180}{9} = 20 \qquad \bar{Y} = \frac{\Sigma Y}{N} = \frac{360}{9} = 40$$

byx = Coefficient of regression of Y on X $= \dfrac{\Sigma xy}{\Sigma x^2} = \dfrac{193}{120} = 1.6083$

bxy = Coefficient of regression of X on Y $= \dfrac{\Sigma xy}{\Sigma y^2} = \dfrac{193}{346} = 0.5578$

Karl Pearson's correlation coefficient r between x and y is given by:

$$r^2 = byx \,.\, bxy = 1.6083 \times 0.5578 = 0.8971$$

$$\Rightarrow \quad r = \pm\sqrt{0.8971} = \pm 0.9471$$

Since, the regression coefficients are positive, r is also positive.

$$\therefore \quad r = +0.9471$$

Aliter:

$$r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \,.\, \Sigma y^2}} = \frac{193}{\sqrt{120 \times 346}} = \frac{193}{\sqrt{41520}}$$

$$= \frac{193}{203.7646} = 0.9471$$

Thus, we see that there is a very high degree of positive correlation between the test scores (X) and the sales ('00 Rs.) (Y). This justifies the proposal for the termination of service of those with low test scores.

Regression Equations

To obtain the test score (X) for given sales (Y), we use the equation of the line of regression of X on Y.

The equation of line of regression of X on Y is:

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$

$$X - 20 = 0.5578\,(Y - 40) = 0.5578Y - 22.312$$

$$\Rightarrow \quad X = 0.5578Y - 22.312 + 20$$

$$\Rightarrow \quad \boxed{X = 0.5578Y - 2.312} \qquad \ldots(i)$$

Hence, to ensure the continuation of service, the minimum test score (X) corresponding to a minimum sales volume (Y) of Rs. 3,000 = 30 ('00 Rs.) is obtained on putting $Y = 30$ in (i) and is given by :

$$X = 0.5578 \times 30 - 2.312 = 16.734 - 2.312$$

$$= 14.422 \approx 14$$

To estimate the sales volume (Y) of a salesman with given test score (X), we use the line of regression of Y on X, which is given by :

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 40 = 1.6083\,(X - 20) = 1.6083X - 32.1660$$

$$\Rightarrow \quad Y = 1.6083X - 32.1660 + 40$$

$$\Rightarrow \quad \boxed{Y = 1.6083X + 7.8340}$$

Hence, the estimated sales volume of a salesman with test score of 28 is (in '00 Rs.)

$$Y = 1.6083 \times 28 + 7.8340$$

$$= 45.0324 + 7.8340$$

$$= 52.8664 \text{ ('00 Rs.)} = \text{Rs. } 5286.64$$

Example 58. In the estimation of regression equation of two variables X and Y, the following results were obtained: $\bar{X} = 90, \bar{Y} = 70, N = 10, \Sigma x^2 = 6360, \Sigma y^2 = 2860, \Sigma xy = 3900$, where, x and y are deviations from their respective means. Obtain the two lines of regression.

Solution: Given, $N = 10, \bar{X} = 90, \bar{Y} = 70, \Sigma x^2 = 6360, \Sigma y^2 = 2860, \Sigma xy = 3900$

$$\therefore \quad byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{3900}{6360} = +0.61 \quad \text{Where, } x = X - \bar{X};\ y = Y - \bar{Y}$$

$$bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{3900}{2860} = 1.36$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$Y - 70 = 0.61\,(X - 90)$$

$$Y - 70 = 0.61X - 54.9$$
$$Y = 0.61X + 15.1$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 90 = 1.36\,(Y - 70)$$
$$X - 90 = 1.36Y - 95.2$$
$$X = 1.36Y - 5.2$$

Example 59. From the following data, find out (*i*) correlation coefficient, (*ii*) linear regression line of Y on X. Also find out the percentage of variation explained by the regression line.

X :	1	2	3	4	5
Y :	2	5	3	8	7

Solution :

Calculations of Correlation Coefficient and Regression Equations

X	$\bar{X}=3$ x	x^2	Y	$\bar{Y}=5$ y	y^2	xy
1	−2	4	2	−3	9	6
2	−1	1	5	0	0	0
3	0	0	3	−2	4	0
4	+1	1	8	+3	9	3
5	+2	4	7	+2	4	4
$\Sigma X = 15$ $N = 5$	$\Sigma x = 0$	$\Sigma x^2 = 10$	$\Sigma Y = 25$	$\Sigma y = 0$	$\Sigma y^2 = 26$	$\Sigma xy = 13$

$$\bar{X} = \frac{15}{5} = 3, \bar{Y} = \frac{25}{5} = 5$$

Since the actual means of X and Y are whole numbers, we should take deviations from actual means of X and Y to simplify the calculations.

(*i*) Calculation of Correlation Coefficient

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{13}{\sqrt{10 \times 26}} = \frac{13}{16.12} = 0.806$$

(*ii*) Calculation of Regression Equations

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{13}{10} = 1.3$$

$$bxy = \frac{\Sigma xy}{\Sigma y^2} = \frac{13}{26} = 0.5$$

Regression Equation of Y on X

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 5 = 1.3\,(X - 3)$$
$$Y - 5 = 1.3X - 3.9 \quad \Rightarrow \quad Y = 1.3X + 1.1$$

Regression Equation of X on Y

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 3 = 0.5\,(Y - 5)$$
$$X - 3 = 0.5Y - 2.5 \quad \Rightarrow \quad X = 0.5Y + 0.5$$

Calculation of r^2

r^2 = Coefficient of Determination
$= (0.806)^2 = 0.6496 = 64.96\%$

This implies that 64.96% variations in Y are explained by the regression line of Y on X.

Example 60. Given the regression equation of Y on X and X on Y are respectively. $Y = 2X$ and $6X - Y = 4$ and the second moment of X about origin $\left(i.e., \Sigma X^2/N\right)$ is 3. (*i*) Find the correlation coefficient and (*ii*) standard deviation of y.

Solution: (*i*) Regression equation of Y on X is $Y = 2X$ $\Rightarrow$ $b_{yx} = 2$

Regression equation of X on Y is $6X - Y = 4$ $\Rightarrow$ $X = \frac{1}{6}Y + \frac{4}{6} \Rightarrow b_{xy} = \frac{1}{6}$

$$\therefore \quad r = \sqrt{2 \times \frac{1}{6}} = \sqrt{0.3333} = 0.578$$

(*ii*) Second moment of X about origin = 3 (given)

$$\Rightarrow \quad \frac{\Sigma X^2}{N} = 3$$

$$\sigma_x^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = 3 - (\bar{X})^2$$

Solving the two regression equations, we get $\bar{X} = 1, \bar{Y} = 2$

$$\therefore \quad \sigma_x^2 = 3 - 1 = 2 \quad \therefore \quad \sigma_x = \sqrt{2} = 1.414$$

$$byx = r\frac{\sigma_y}{\sigma_x}$$

i.e., $\quad 2 = 0.578 \cdot \frac{\sigma_y}{\sqrt{2}}$ *i.e.*, $\quad \sigma_y = \frac{2\sqrt{2}}{0.578} = \frac{2.828}{0.578} = 4.89$

Example 61. Find the 'standard error of the estimates'

$$\sigma_x = 1.414,\ \sigma_y = 8.94,\ r = 0.316$$

Solution: Given, $r = 0.316, \sigma_x = 1.414, \sigma_y = 8.94$

(*i*) $S_{xy} = \sigma_x \cdot \sqrt{1 - r^2} = 1.414\sqrt{1 - (0.316)^2} = 1.414 \times 0.949 = 1.342$

(*ii*) $S_{yx} = \sigma_y \cdot \sqrt{1 - r^2} = 0.894\sqrt{1 - (0.316)^2} = 0.894 \times 0.949 = 0.848$

Example 62. For a set of 10 pairs of values of X and Y, the regression line of X on Y is $X - 2Y + 12 = 0$. The mean and standard deviation of Y being 8 and 2 respectively. Later it is known that a pair ($X = 3$, $Y = 8$) was wrongly recorded and the correct pair detected is ($X = 8$, $Y = 3$). Find the correct regression line of X on Y.

Solution: In the usual notations we are given, $N = 10$, $\bar{Y} = 8$, $\sigma_y = 2$

The equation of the line of regression of X on Y is : $X - 2Y + 12 = 0$ (given). Since the lines of regression pass through the point $(\bar{X}, \bar{Y})$, we get

$$\bar{X} - 2\bar{Y} + 12 = 0 \quad \Rightarrow \quad \bar{X} = 2\bar{Y} - 12 = 2 \times 8 - 12 = 4$$

$$X - 2Y + 12 = 0 \quad \Rightarrow \quad X = 2Y - 12 \quad \Rightarrow \quad b_{xy} = 2 \quad [\text{Using (i)}]$$

$$b_{xy} = \frac{\frac{\Sigma XY}{N} - \bar{X}\bar{Y}}{\sigma_y^2} = 2 \quad \Rightarrow \quad \frac{\Sigma XY}{N} - \bar{X}\bar{Y} = 2 \times 2^2 = 8$$

$$\Rightarrow \quad \frac{\Sigma XY}{N} - \bar{X}\bar{Y} = 8 \quad \Rightarrow \quad \Sigma XY = 10[8 + 4 \times 8] = 10 \times 40 = 400$$

$$\sigma_y^2 = \frac{\Sigma Y^2}{N} - (\bar{Y})^2 \quad \Rightarrow \quad \Sigma Y^2 = N[\sigma_y^2 + \bar{Y}^2] = 10\,[4 + 8^2] = 680$$

We have $\bar{X} = 4$, $\bar{Y} = 8$, $\Sigma Y^2 = 680$, $\Sigma XY = 400$

Wrong pair = (X = 3, Y = 8); Correct pair = (X = 8, Y = 3)

Calculation of Correct Values

$$\bar{X} = \frac{\Sigma X}{N} \quad \Rightarrow \quad 4 = \frac{\Sigma X}{10} \quad \Rightarrow \quad \Sigma X = 40$$

Corrected $\Sigma X = 40 - 3 + 8 = 45$ $\Rightarrow$ Corrected $\bar{X} = \frac{45}{10} = 4.5$

$$\bar{Y} = \frac{\Sigma Y}{N} \quad \Rightarrow \quad 8 = \frac{\Sigma Y}{10} \quad \Rightarrow \quad \Sigma Y = 80$$

Corrected $\Sigma Y = 80 - 8 + 3 = 75$ $\Rightarrow$ Corrected $\bar{Y} = \frac{75}{10} = 7.5$

Corrected $\Sigma Y^2 = 680 - 8^2 + 3^2 = 625$

Corrected $\sigma_y^2 = \frac{\Sigma Y^2}{N} - (\bar{Y})^2 = \frac{625}{10} - (7.5)^2 = 6.25$

Corrected $\Sigma XY = 400 - 24 + 24 = 400$

$$\text{Corrected } b_{xy} = \frac{\frac{\Sigma XY}{N} - \bar{X}\bar{Y}}{\sigma_y^2} = \frac{\frac{400}{10} - (4.5)(7.5)}{6.25} = 1$$

Corrected line of regression of X on Y becomes

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

$$\Rightarrow \quad X - 4.5 = 1(Y - 7.5)$$

$$\Rightarrow \quad X = Y - 7.5 + 4.5 \quad \Rightarrow \quad X = Y - 3$$

Example 63. The data in the following table relates the weekly maintenance cost (in Rs.) to the age (in months) of ten machines of similar type in a manufacturing company. Find the least squares regression line of maintenance cost on age and use to predict the maintenance cost for a machine of this type which is 40 months old.

Machine:	1	2	3	4	5	6	7	8	9	10
Age (X):	5	10	15	20	30	30	30	50	50	60
Cost (Y):	190	240	250	300	310	335	300	300	350	395

Solution:

Computation of Regression Equation

X	Y	$X - \bar{X} = x$	$Y - \bar{Y} = y$	x^2	xy
5	190	− 25	− 107	625	2,675
10	240	− 20	− 57	400	1,140
15	250	− 15	− 47	225	705
20	300	− 10	3	100	− 30
30	310	0	13	0	0
30	335	0	38	0	0
30	300	0	3	0	0
50	300	20	3	400	60
50	350	20	53	400	1,060
60	395	30	98	900	2,940
$\Sigma X = 300$	$\Sigma Y = 2,970$	$\Sigma x = 0$	$\Sigma y = 0$	$\Sigma x^2 = 3,050$	$\Sigma xy = 8,550$

$$\bar{X} = \frac{300}{10} = 30; \quad \bar{Y} = \frac{2970}{10} = 297$$

Since both means are integers deviations have been taken from actual means.

Regression Coefficient of Y on X

$$byx = \frac{\Sigma xy}{\Sigma x^2} = \frac{8,550}{3,050} = 2.8033$$

Regression Equation of Y on X is

$$Y - \bar{Y} = byx\,(X - \bar{X})$$

$$\Rightarrow \quad y - 297 = 2.8033\,(X - 30) = 2.8033X - 84.099$$

$$\Rightarrow \quad Y = 2.8033X + 212.901$$

For $X = 40$, $Y = 2.8033 \times 40 + 212.901$

$= 325.033 = \text{Rs. } 325.$

Example 64. For certain data, the following regression equations were obtained :

$$4X - 5Y + 33 = 0$$

$$20X - 9Y - 107 = 0$$

Estimate Y when $X = 20$ and X when $Y = 20$.

Solution : Let $4X - 5Y + 33 = 0$ be the regression equation of Y on X, while $20X - 9Y - 107 = 0$ be the regression equation of X on Y,

From (*i*) $\quad Y = \frac{4}{5}X + \frac{33}{5} \Rightarrow \quad byx = \frac{4}{5}$ and

From (*ii*) $\quad X = \frac{9}{20}Y + \frac{107}{20} \Rightarrow \quad bxy = \frac{9}{20}$

Now, $\quad r = \sqrt{byx \times bxy} = \sqrt{\frac{4}{5} \times \frac{9}{20}} = \sqrt{\frac{9}{25}} = \frac{3}{5} < 1,$

So our assumption is correct.

$\therefore$ When $X = 20$, $Y_{20} = \frac{4}{5} \times 20 + \frac{33}{5} = \frac{113}{5} = 22.6$ and

When $Y = 20$, $X_{20} = \frac{9}{20} \times 20 + \frac{107}{20} = \frac{287}{20} = 14.35.$

Example 65. Given the following data, find what will be (a) the height of a policeman whose weight is 200 pounds, (b) the weight of a policeman who is 6 ft. tall.
Average height = 68 inches, average weight = 150 pounds, coefficient of correlation between height and weight = 0.6, S.D. of heights = 2.5 inches, S.D. of weights = 20 pounds.

Solution: Let height of policeman be denoted by X and weight of policeman by Y.

We are given:

$\bar{X} = 68''$, $\quad \bar{Y} = 150$ lbs, $\quad \sigma_x = 2.5''$, $\quad \sigma_y = 20$ lbs, $\quad r_{xy} = 0.6$

(*i*) For estimating the height of a policeman whose weight is 200 lbs, we use regression of X on Y as follows:

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

$$X - 68 = 0.6 \times \frac{2.5}{20}(Y - 150)$$

$$X - 68 = 0.075\,(Y - 150)$$

$$X - 68 = 0.075Y - 11.25$$

$$X = 0.075Y + 56.75$$

When $Y = 200$, X = 0.075 (200) + 56.75 = 71.75

Thus, the height of a policeman whose weight is 200 lbs shall be 71.75".

(*ii*) For estimating the weight of a policeman whose height is 72" (*i.e.*, 6 ft), we use regression of Y on X as follows:

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

$$Y - 150 = 0.6 \times \frac{20}{2.5}(X - 68)$$

$$Y - 150 = 4.8\,(X - 68)$$

$$Y - 150 = 4.8X - 326.4$$

$$Y = 4.8X - 176.4$$

When $X = 72$, $Y = 4.8\,(72) - 176.4 = 169.2$

Thus, the weight of a policeman who is 6 ft. tall should be 169.2 lbs.

Example 66. Prove that regression coefficients are independent of the change of origin but not of scale.

Solution: Change the X and Y variables into new variables in the following manner:

$$U = \frac{x - a}{h},\ V = \frac{y - b}{k}$$

$$bxy = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(Y - \bar{Y})^2} \qquad \ldots(i)$$

$X = hu + a$ $\qquad$ $Y = kv + b$

$\Sigma X = h\Sigma u + \Sigma a$ $\qquad$ $\Sigma Y = k\Sigma v + \Sigma b$

$\Sigma X = h\Sigma u + na$ $\qquad$ $\Sigma Y = k \cdot \Sigma v + nb$

$\frac{\Sigma X}{n} = h \cdot \frac{\Sigma u}{n} + a$ $\qquad$ $\frac{\Sigma Y}{n} = k \cdot \frac{\Sigma v}{n} + b$

$\therefore \quad \bar{X} = h\bar{u} + a \qquad \therefore \quad \bar{Y} = k\bar{v} + b$

$\therefore \quad X - \bar{X} = h(u - \bar{u}) \qquad \therefore \quad Y - \bar{Y} = k(v - \bar{v})$

Substituting in (*i*),

$$bxy = \frac{\Sigma h(u - \bar{u}) \cdot k \cdot (v - \bar{v})}{\Sigma k^2 (v - \bar{v})^2} = \frac{hk}{k^2} \cdot \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\Sigma(v - \bar{v})^2}$$

$$bxy = \frac{h}{k} \cdot \frac{\Sigma(u - \bar{u})(v - \bar{v})}{\Sigma(v - \bar{v})^2} = \frac{h}{k} \cdot buv$$

Hence the result is proved.

Similarly, we can prove for $byx = \frac{k}{h}bvu$

Example 67. Prove that the mean of two regression coefficients is always greater than the coefficient of correlation.

Solution: We have to prove

$$\frac{byx + bxy}{2} > r$$

$$\frac{r \cdot \frac{\sigma_y}{\sigma_x} + r \cdot \frac{\sigma_x}{\sigma_y}}{2} > r \qquad \Rightarrow \qquad \frac{r\left(\frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y}\right)}{2} > r$$

$$\text{or} \quad \frac{\frac{\sigma_y}{\sigma_x}+\frac{\sigma_x}{\sigma_y}}{\frac{\sigma_x}{\sigma_y}-\frac{\sigma_y}{\sigma_x}} > 1 \quad \Rightarrow \quad \frac{\sigma^2_y+\sigma^2_x}{2\sigma_x\cdot\sigma_y} > 1$$

Multiplying both sides by $2\sigma_x.\sigma_y$

$$\Rightarrow \quad \sigma^2_x+\sigma^2_y > 2\sigma_x\cdot\sigma_y$$

$$\Rightarrow \quad \sigma^2_x+\sigma^2_y-2\sigma_x.\sigma_y \geq 0 \quad \Rightarrow \quad (\sigma_x-\sigma_y)^2 \geq 0 \qquad [\because \ \sigma_x>0, \sigma_y>0]$$

As the square of real numbers can never be less than zero. Hence the arithmetic mean of the two regression coefficients is greater than the correlation coefficient.

Example 68. Show that θ, the acute angle between two lines of regression is given by:

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

Interpret the case when $r = 0, \pm 1$

Solution: Equations of the two lines of regression are:

$$Y-\bar{Y} = r.\frac{\sigma_y}{\sigma_x}(X-\bar{X}) \text{ and } X-\bar{X} = r.\frac{\sigma_x}{\sigma_y}(Y-\bar{Y})$$

We have, m_1 = slope of the line of regression of Y on X = $byx = r.\frac{\sigma_y}{\sigma_x}$

m_2 = slope of the line of regression of X on Y = $\frac{1}{bxy} = \frac{1}{r}.\frac{\sigma_y}{\sigma_x}$

Let θ be the angle between two lines of regression, then

$$\tan\theta = \pm\frac{m_2-m_1}{1+m_1m_2} = \frac{\frac{1}{r}.\frac{\sigma_y}{\sigma_x}-r.\frac{\sigma_y}{\sigma_x}}{1+r.\frac{\sigma_y}{\sigma_x}.\frac{1}{r}.\frac{\sigma_y}{\sigma_x}} = \pm\left(\frac{1-r^2}{r}\right)\left(\frac{\sigma_x\cdot\sigma_y}{\sigma_x^2+\sigma_y^2}\right)$$

Since $r^2 \leq 1$ and σ_x, σ_y are positive.

∴ +ve sign gives the acute angle between the lines.

Hence, $\tan\theta = \frac{(1-r^2)}{|r|}.\frac{\sigma_x\cdot\sigma_y}{\sigma_x^2+\sigma_y^2}$

Case I: When $r = 0, \tan\theta = \infty \Rightarrow \theta = \frac{\pi}{2} = 90°$

Thus, the lines of regression are perpendicular to each other.

Case II: When $r = \pm 1, \tan\theta = 0 \Rightarrow \theta = 0$ or π

Thus, the two lines of regression coincide and there will be one regression line.

IMPORTANT FORMULAE

1. Regression Lines

(i) Regression Equation of Y on X

$$Y-\bar{Y} = byx(X-\bar{X})$$

Where, $byx = \frac{N.\Sigma XY-\Sigma X.\Sigma Y}{N.\Sigma X^2-(\Sigma X)^2}$ (When we use actual values of X and Y)

$= \frac{\Sigma xy}{\Sigma x^2}$ Where, $x = X-\bar{X}$; $Y = Y-\bar{Y}$

$= \frac{N\cdot\Sigma dxdy-\Sigma dx\cdot\Sigma dy}{N\cdot\Sigma dx^2-(\Sigma dx)^2}$ (When deviations are taken from actual means of X and Y)

Where, $dx = X-A$, $dy = Y-A$

$= r.\frac{\sigma_y}{\sigma_x}$ (When deviations are taken from assumed means of X and Y)

(When we use r, σ_y and σ_x)

(ii) Regression Equation of X on Y

$$X-\bar{X} = bxy(Y-\bar{Y})$$

$bxy = \frac{N.\Sigma XY-\Sigma X.\Sigma Y}{N.\Sigma Y^2-(\Sigma Y)^2}$ (When we use actual values of X and Y)

$= \frac{\Sigma xy}{\Sigma y^2}$ (When deviations are taken from actual mean)

$= \frac{N\cdot\Sigma dxdy-\Sigma dx\cdot\Sigma dy}{N\cdot\Sigma dy^2-(\Sigma dy)^2}$ (When deviations are taken from assumed mean)

$= r.\frac{\sigma_x}{\sigma_y}$ (When we use r, σ_x and σ_y)

(iii) In Grouped Frequency Distribution

$$byx = \frac{N.\Sigma fdxdy-\Sigma fdx\,\Sigma fdy}{N.\Sigma fdx^2-(\Sigma fdx)^2}\times\frac{i_y}{i_x}$$

and $$bxy = \frac{N.\Sigma fdxdy-\Sigma fdx\,\Sigma fdy}{N.\Sigma fdy^2-(\Sigma fdy)^2}\times\frac{i_x}{i_y}$$

Where, $N = \Sigma f$, stands for the total frequency.

2. Regression Coefficients

There are two regression coefficients:

(*i*) Regression coefficient of Y on X = $byx = r \cdot \frac{\sigma_y}{\sigma_x}$

(*ii*) Regression coefficient of X on Y = $bxy = r \cdot \frac{\sigma_x}{\sigma_y}$

(*iii*) $r = \sqrt{bxy \times byx}$

3. Standard Error

Standard error of estimate $S_{xy} = \sqrt{\frac{\Sigma(X - X_c)^2}{N}}$ or $S_{xy} = \sigma_x \cdot \sqrt{1 - r^2}$

$S_{yx} = \sqrt{\frac{\Sigma(Y - Y_c)^2}{N}}$ or $S_{yx} = \sigma_y \cdot \sqrt{1 - r^2}$

Important Note

(*i*) Regression equation of Y on X is used to estimate the best (average) value of Y for given value of X.

(*ii*) Regression equation of X on Y is used to estimate the best (average) value of X for given value of Y.

QUESTIONS

1. Explain the concept of regression and comment on its utility. Also distinguish between correlation and regression.
2. What are regression coefficients? Explain the properties of regression coefficients.
3. Discuss the difference between correlation and regression.
4. Define the Standard Error of Estimate. How is it computed?
5. What is regression line? Why are there, in general, two regression lines? Under what conditions can there be only one regression line? When the two lines of regression intersect each other at 90°?
6. What would be lines of regression if $r = +1, r = -1, r = 0$. Give interpretation in each case.
7. Explain the meaning of (*i*) Standard Error of Estimate and (*ii*) Coefficient of Determination.
8. How would you identify regression equation of X on Y and Y on X?
9. What is the relationship between correlation and regression coefficients?

Index Numbers–I

3

INTRODUCTION

Economic and business data change from time to time. For instance, prices of all commodities do not remain constant. It is possible that sometimes the price of a commodity rises and sometimes falls. Similarly, output of a commodity sometimes rises and sometimes falls. The measurement of such changes is possible only by means of some statistical methods. Index numbers are such statistical devices which help in the measurement of such changes. In other words, by index numbers we mean the statistical measures with the help of which relative changes in general price level taking place at different points of time can be measured. Application of Index numbers are not limited only to general price levels but rather they help in the relative measurement of every such phenomenon like cost-of-living, output, national income, business activities whose direct measurement is not possible. Index numbers are used to measure the relative changes in some phenomena which we cannot observe directly.

DEFINITION OF INDEX NUMBERS

Some important definitions of index numbers are given below:

1. Index Numbers are a specialized type of averages. —M. Blair
2. Index Numbers are devices for measuring differences in the magnitude of a group of related variables. —Croxten and Cowden
3. An Index Number is a statistical measure designed to show changes in a variable or group of related variables with respect to time, geographic location or other characteristics. —Spiegel

The definitions discussed above specify the following features of index numbers:

(*i*) Relative changes in the aggregates are measured by index numbers (*ii*) Index numbers always present the changes taking place in some variable on an average only (*iii*) By index numbers, positions in base year and current year are compared.

USES OF INDEX NUMBERS

In present times, the importance of index numbers is increasing. Nowadays, they are being used in economics and business fields. To quot **Simpson and Kafka, "Index Numbers are economic barometers".** The main uses of index numbers are the followings:

(I) **To Simplify Complexities:** An index number makes possible the measurement of such complex changes whose direct measurement is not possible. In other words, index numbers are used to measure the changes in some quantity which we cannot observe directly.