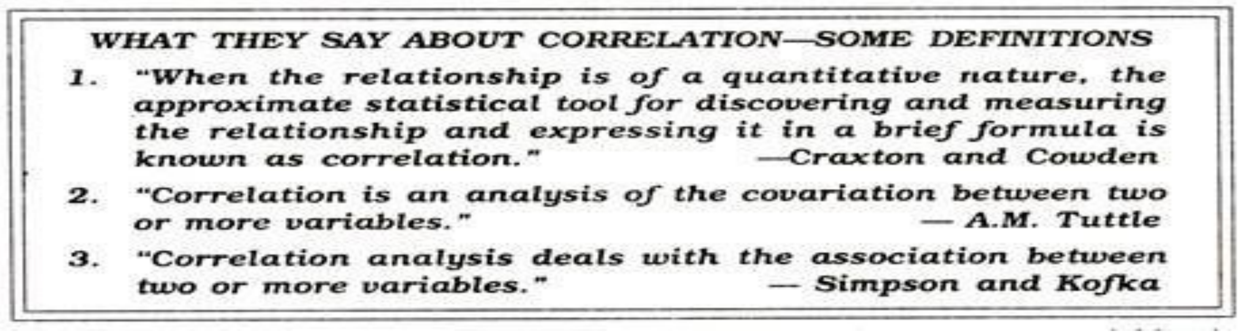
CORRELATION

Definition: The **Correlation** is a statistical tool used to measure the relationship between two or more variables, i.e. the degree to which the variables are associated with each other, such that the change in one is accompanied by the change in another.



WHAT THEY SAY ABOUT CORRELATION—SOME DEFINITIONS

1. *"When the relationship is of a quantitative nature, the approximate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation."* —Craxton and Cowden
2. *"Correlation is an analysis of the covariation between two or more variables."* — A.M. Tuttle
3. *"Correlation analysis deals with the association between two or more variables."* — Simpson and Kofka

Types of Correlation:

1. Positive Correlation

A correlation in the same direction is called a positive correlation. If one variable increases the other also increases and when one variable decreases the other also decreases. For example, the length of an iron bar will increase as the temperature increases.

- Price and Supply
- Sales and Expenditure on Advertisement
- Yield and Fertilizer Applied

2. Negative Correlation

Correlation in the opposite direction is called a negative correlation. Here if one variable increases the other decreases and vice versa.

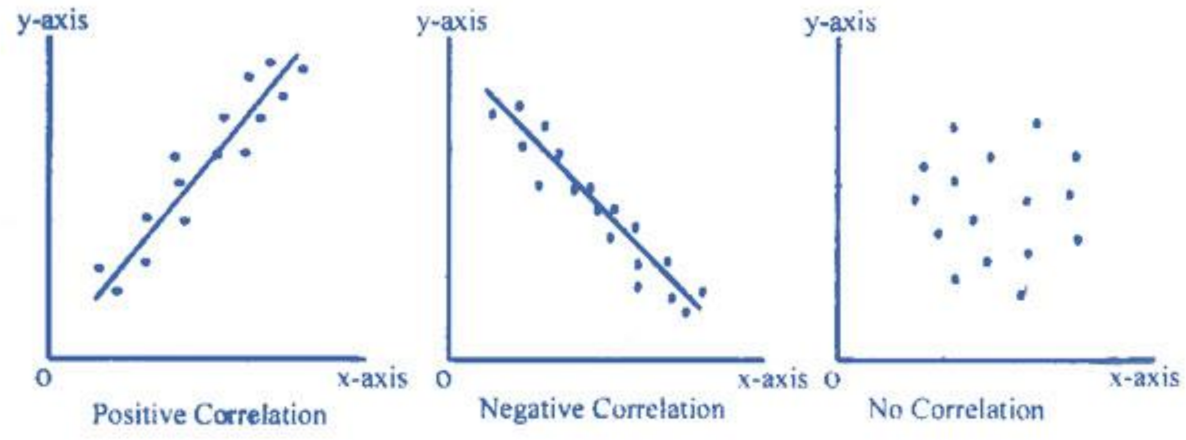
For example, the volume of gas will decrease as the pressure increases, or the demand for a particular commodity increases as the price of such commodity decreases.

Examples:

- Price and Demand
- Yield and Weed

3. No Correlation or Zero Correlation

If there is no relationship between the two variables such that the value of one variable changes and the other variable remains constant, it is called no or zero correlation.



4.Simple Correlation

When only two variables are considered as under positive or negative correlation, the correlation between them is called simple correlation.

5.PartialCorrelation

When more than two variables are considered, the correlation between two of them when all other variables are held constant, i.e. when the linear effects of all other variables on them are removed , is called partial correlation.

6.Multiple Correlation

When more than two variables are considered, the correlation between one of them and its estimate based on the group consisting of the other variables is called partial correlation.

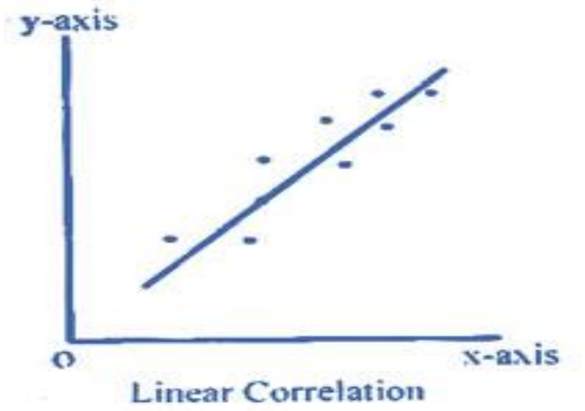
7.Linear Correlation

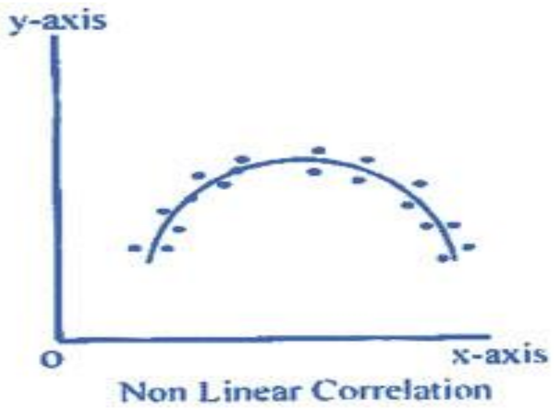
Correlation is said to be linear if the ratio of change is constant. When the amount of output in a factory is doubled by doubling the number of workers, this is an example of linear correlation.

In other words, when all the points on the scatter diagram tend to lie near a line which looks like a straight line, the correlation is said to be linear. This is shown in the figure on the left below

8.Non Linear (Curvilinear) Correlation

Correlation is said to be non linear if the ratio of change is not constant. In other words, when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is said to be non linear (curvilinear). This is shown in the figure on the right below.





Methods of Determining Correlation:

Scatter Diagram Method.
Karl Pearson's Coefficient of Correlation.
Spearman's Rank Correlation Coefficient; and.
Methods of Least Squares.

Karl Pearson's Coefficient of Correlation
Spearman's Rank Correlation Coefficient
Method of Least Squares
Scatter Diagram Method
Correlation Methods

Scatter Diagram Method

Definition: The Scatter Diagram Method is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the number of observations. Then by looking at the scatter of several points, the degree of correlation is ascertained.

The degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart. The more the points plotted are scattered over the chart, the lesser is the degree of correlation between the variables. The more the points plotted are closer to the line, the higher is the degree of correlation. The degree of correlation is denoted by **“r”.**

The following types of scatter diagrams tell about the degree of correlation between variable X and variable Y.

Perfect Positive Correlation (r=+1): The correlation is said to be perfectly positive when all the points lie on the straight line rising from the lower left-hand corner to the upper right-hand corner.

Perfect Negative Correlation (r=-1): When all the points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner, the variables are said to be negatively correlated.

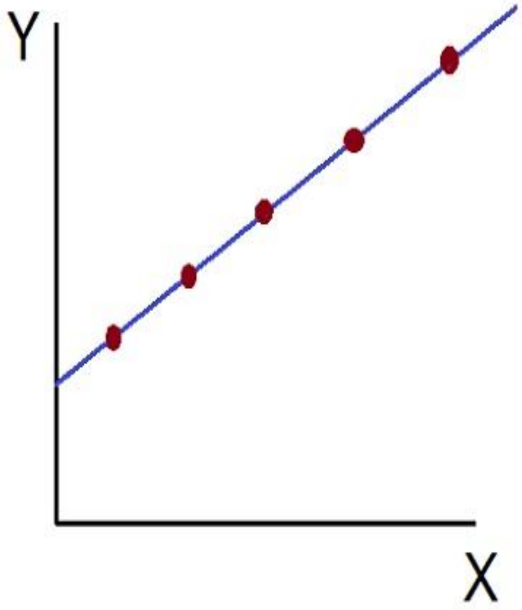
High Degree of +Ve Correlation (r= + High): The degree of correlation is high when the points plotted fall under the narrow band and is said to be positive when these show the rising tendency from the lower left-hand corner to the upper right-hand corner.

High Degree of –Ve Correlation (r= – High): The degree of negative correlation is high when the point plotted fall in the narrow band and show the declining tendency from the upper left-hand corner to the lower right-hand corner.

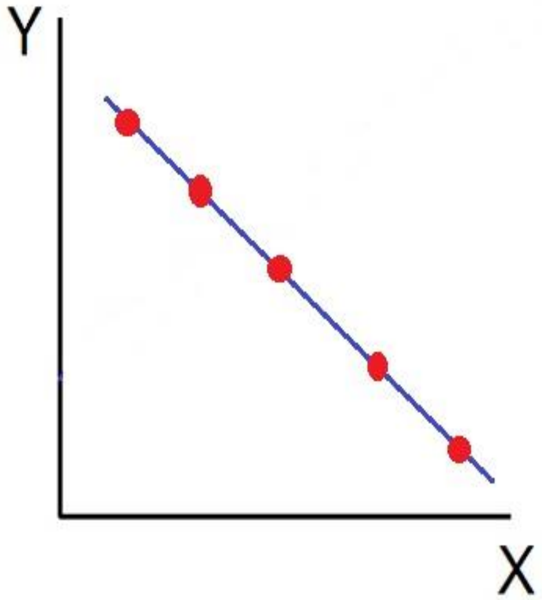
Low degree of +Ve Correlation (r= + Low): The correlation between the variables is said to be low but positive when the points are highly scattered over the graph and show a rising tendency from the lower left-hand corner to the upper right-hand corner.

Low Degree of –Ve Correlation (r= + Low): The degree of correlation is low and negative when the points are scattered over the graph and the show the falling tendency from the upper left-hand corner to the lower right-hand corner.

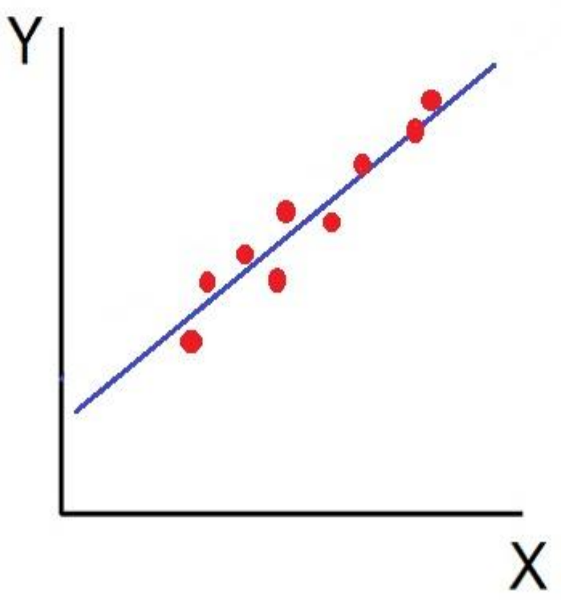
No Correlation (r= 0): The variable is said to be unrelated when the points are haphazardly scattered over the graph and do not show any specific pattern. Here the correlation is absent and hence **r = 0**.

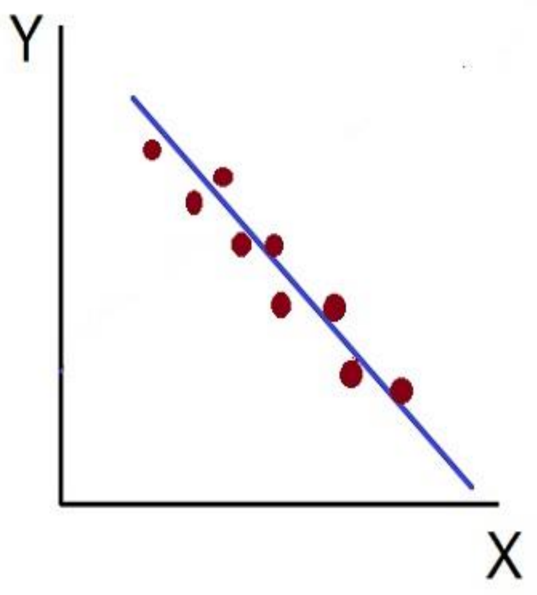


PerfectPositiveCorrelation



Perfect Negative Correlation





High Degree of +Ve Correlation **High Degree of –Ve Correlation**

Thus, the scatter diagram method is the simplest device to study the degree of relationship between the variables by plotting the dots for each pair of variable values given. The chart on which the dots are plotted is also called as a **Dotogram**.

Karl Pearson's Coefficient of Correlation

Definition: **Karl Pearson's Coefficient of Correlation** is widely used mathematical method wherein the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Pearson's method, popularly known as a **Pearsonian Coefficient of Correlation,** is the most extensively used quantitative methods in practice. The coefficient of correlation is denoted by **"r".**

Formula:

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

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

Properties of Coefficient of Correlation:

- The value of the coefficient of correlation (r) always **lies between ±1**. Such as:
  r=+1,perfectpositivecorrelation
  r=-1,perfectnegativecorrelation
  r=0, no correlation
- The coefficient of correlation is independent of the **origin and scale.** By origin, it means subtracting any non-zero constant from the given value of X and Y the value of "r" remains unchanged. By scale it means, there is no effect on the value of "r" if the value of X and Y is divided or multiplied by any constant.
  - The coefficient of correlation is a **geometric mean of two regression coefficient.** Symbolically it is represented as:

$$\mathrm{r} = \sqrt{b_{xy} + b_{yx}}$$

- The coefficient of correlation is **" zero"** when the variables X and Y are independent. But, however, the converse is not true.

Assumptions of Karl Pearson's Coefficient of Correlation:

1. The relationship between the variables is **"Linear",** which means when the two variables are
plotted, a straight line is formed by the points plotted.
2. There are a large number of independent causes that affect the variables under study so as to
form a **Normal Distribution**. Such as, variables like price, demand, supply, etc. are affected by such factors that the normal distribution is formed.
3. The variables are independent of each other.

Note: The coefficient of correlation measures not only the magnitude of correlation but also tells the direction. Such as, r = -0.67, which shows correlation is negative because the sign is **"-"** and the magnitude is **0.67**.

Spearman's Rank Correlation Coefficient:

Definition: The **Spearman's Rank Correlation Coefficient** is the non-parametric statistical measure used to study the strength of association between the two ranked variables. This method is applied to the ordinal set of numbers, which can be arranged in order, i.e. one after the other so that ranks can be given to each.

In the rank correlation coefficient method, the ranks are given to each individual on the basis of its quality or quantity, such as ranking starts from position 1st and goes till Nth position for the one ranked last in the group.

The formula to calculate the rank correlation coefficient is:

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

Where,ρ=Rankcoefficientofcorrelation

d=Differenceofranks

N = Number of Observations

The value of ρ lies between ±1 such as:

ρ =+1, there is a complete agreement in the order of ranks and move in the same direction.

ρ=-1, there is a complete agreement in the order of ranks, but are in opposite directions.

ρ =0, there is no association in the ranks.

While solving for the rank correlation coefficient one may come across the following problems:

- Where actual Ranks are given
- Where ranks are not given
- Equal Ranks or Tie in Ranks

Where actual ranks are given: An individual must follow the following steps to calculate the correlation coefficient:

1. First, the difference between the ranks (R1-R2) must be calculated, denoted by D.
2. Then, square these differences to remove the negative sign and obtain its sum $\sum D^2$.
3. Apply the formula as shown above.

Where ranks are not given: In case the ranks are not given, then the individual may assign the rank by taking either the highest value or the lowest value as 1. Whatever criteria is being decided the same method should be applied to all the variables.

Equal Ranks or Tie in Ranks: In case the same ranks are assigned to two or more entities, then the ranks are assigned on an average basis. Such as if two individuals are ranked equal at third position, then the ranks shall be calculated as:(3+4)/2 = 3.5

The formula to calculate the rank correlation coefficient when there is a tie in the ranks is:

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

Where m = number of items whose ranks are common.

Note: The Spearman's rank correlation coefficient method is applied only when the initial data are in the form of ranks, and N (number of observations) is fairly small, i.e. not greater than 25 or 30.

Method of Least Squares:

Definition: The **Method of Least Squares** is another mathematical method that tells the degree of correlation between the variables by using the square root of the product of two regression coefficient that of x on y and y on x.

Coefficient of Determination:

Definition: The **Coefficient of determination** is the square of the coefficient of correlation r^2 which is calculated to interpret the value of the correlation. It is useful because it explains the level of variance in the dependent variable caused or explained by its relationship with the independent variable.

The coefficient of determination explains the proportion of the explained variation or the relative reduction in variance corresponding to the regression equation rather than about the mean of the dependent variable. For example, if the value of $r = 0.8$, then r^2 will be 0.64, which means that 64% of the variation in the dependent variable is explained by the independent variable while 36% remains unexplained.

Thus, the coefficient of determination is the ratio of explained variance to the total variance that tells about the strength of linear association between the variables, say X and Y. The value of r^2 lies between 0 and 1 and observes the following relationship with 'r'.

- With the decrease in the value of 'r' from its maximum value of 1, the 'r^2' also decreases much more rapidly.
- The value of 'r' will always be greater than 'r^2' unless the $r^2 =0$ or 1.

The coefficient of determination also explains that how well the regression line fits the statistical data. The closer the regression line to the points plotted on a scatter diagram, the more likely it explains all the variation and the farther the line from the points the lesser is the ability to explain the variance.

REGRESSION ANALYSIS:

Definition: The Regression Analysis is a statistical tool used to determine the probable change in one variable for the given amount of change in another. This means, the value of the unknown variable can be estimated from the known value of another variable.

The degree to which the variables are correlated to each other depends on the Regression Line. The regression line is a single line that best fits the data, i.e. all the points plotted are connected via a line in the manner that the distance from the line to the points is the smallest.

The regression also tells about the relationship between the two or more variables, then what is the difference between regression and correlation? Well, there are two important points of differences between Correlation and Regression. These are:

- The Correlation Coefficient measures the "degree of relationship" between variables, say X and Y whereas the Regression Analysis studies the "nature of relationship" between the variables.
- Correlation coefficient does not clearly indicate the cause-and-effect relationship between the variables, i.e. it cannot be said with certainty that one variable is the cause, and the other is the effect. Whereas, the Regression Analysis clearly indicates the cause-and-effect relationship between the variables.

The regression analysis is widely used in all the scientific disciplines. In economics, it plays a significant role in measuring or estimating the relationship among the economic variables. For example, the two variables – price (X) and demand (Y) are closely related to each other, so we can find out the probable value of X from the given value of Y and similarly the probable value of Y can be found out from the given value of X.

Regression Line:

Definition: The **Regression Line** is the line that best fits the data, such that the overall distance from the line to the points (variable values) plotted on a graph is the smallest. In other words, a line used to minimize the squared deviations of predictions is called as the regression line.

There are as many numbers of regression lines as variables. Suppose we take two variables, say X and Y, then there will be two regression lines:

- Regression line of Y on X: This gives the most probable values of Y from the given values of X.
- Regression line of X on Y: This gives the most probable values of X from the given values of Y.

The algebraic expression of these regression lines is called as Regression Equations. There will be two regression equations for the two regression lines.

The correlation between the variables depend on the distance between these two regression lines, such as the nearer the regression lines to each other the higher is the degree of correlation, and the farther the regression lines to each other the lesser is the degree of correlation.

The correlation is said to be either perfect positive or perfect negative when the two regression lines coincide, i.e. only one line exists. In case, the variables are independent; then the correlation will be zero, and the lines of regression will be at right angles, i.e. parallel to the X axis and Y axis.

Note: The regression lines cut each other at the point of average of X and Y. This means, from the point where the lines intersect each other the perpendicular is drawn on the X axis we will get the mean value of X. Similarly, if the horizontal line is drawn on the Y axis we will get the mean value of Y.

Regression Equation:

Definition: The **Regression Equation** is the algebraic expression of the regression lines. It is used to predict the values of the dependent variable from the given values of independent variables. If we take two regression lines, say Y on X and X on Y, then there will be two regression equations:

RegressionCoefficientDefinition: The Regression Coefficient is the constant 'b' in the regression equation that tells about the change in the value of dependent variable corresponding to the unit change in the independent variable.

If there are two regression equations, then there will be two regression coefficients:

Methods of Forming The Regression Equations

Method 1: Regression Equations on the basis of Normal Equations

Regression Equation of Y on X: This is used to describe the variations in the value Y from the given changes in the values of X. It can be expressed as follows:

$$Y_e = a + bX$$

Where Y_e is the dependent variable, X is the independent variable and a & b are the two unknown constants that determine the position of the line. The parameter "a" tells about the level of the fitted line, i.e. the distance of a line above or below the origin and parameter "b" tells about the slope of the line, i.e. the change in the value of Y for one unit of change in X.

The values of 'a' and 'b' can be obtained by a method of least squares. According to which the line should be drawn connecting all the plotted points in such a manner that the sum of the squares of the vertical deviations of actual Y from the estimated values of Y is the least, or a best-fitted line is obtained when $\sum (Y-Y_e)^2$ is the minimum.

The following algebraic equations can be solved simultaneously to obtain the values of parameter ‘a’ and ‘b’

$$\sum Y = Na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

Regression Equation of X on Y: This is used to describe the variations in Y from the given changes in the value of X. It can be expressed as follows:

$$X_e = a + bY$$

Where X_e is the dependent variable and Y is the independent variable. The parameters ‘a’ and ‘b’ are the two unknown constants. Again, ‘a’ tells about the level of fitted line and ‘b’ tells about the slope, i.e. the change in the value of X for a unit change in the value of Y.

The following are the two normal equations that can be solved simultaneously to obtain the values of both the parameters ‘a’ and ‘b’.

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Note: The line can be completely determined only if the values of the constant parameters ‘a’ and ‘b’ are obtained.

Method 2: Regression Equations on the basis of $\bar{X}, \bar{Y}, b_{XY}$ and b_{YX}

Regression Equation of Y on X:

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

where $b_{YX} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$ is called regression coefficient of Y on X.

Regression Equation of X on Y:

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

where $b_{XY} = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma Y^2 - (\Sigma Y)^2}$ is called regression coefficient of X on Y.

Note: we use following formulas to find the regression coefficients

1. $b_{YX} = r\frac{\sigma_y}{\sigma_x}$, $b_{YX} = r\frac{\sigma_y}{\sigma_x}$

Here σ_x is standard deviation of X

σ_y is standard deviation of Y

r is correlation coefficient

2. $b_{XY} = \frac{\Sigma xy}{\Sigma y^2}$, $b_{YX} = \frac{\Sigma xy}{\Sigma x^2}$

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

Properties of Regression Coefficient:

Definition: The constant ‘b’ in the regression equation ($Y_e = a + bX$) is called as the **Regression Coefficient**. It determines the slope of the line, i.e. the change in the value of Y corresponding to the unit change in X and therefore, it is also called as a **“Slope Coefficient.”**

Properties:

The correlation coefficient is the **geometric mean** of two regression coefficients. Symbolically, it can be expressed as:

$$\mathbf{r} = \sqrt{\mathbf{b}_{\mathbf{xy}} + \mathbf{b}_{\mathbf{yx}}}$$

The value of the coefficient of correlation **cannot exceed unity i.e. 1.** Therefore, if one of the regression coefficients is greater than unity, the other must be less than unity.

The **sign of both the regression coefficients will be same**, i.e. they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.

The **coefficient of correlation will have the same sign** as that of the regression coefficients, such as if the regression coefficients have a positive sign, then “r” will be positive and vice-versa.

The **average value of the two regression coefficients will be greater than the value of the correlation**. Symbolically, it can be represented as

$$\frac{\mathbf{b}_{\mathbf{xy}} + \mathbf{b}_{\mathbf{yx}}}{\mathbf{2}} > r$$

The regression coefficients are **independent of the change of origin, but not of the scale**. By origin, we mean that there will be no effect on the regression coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or divided by some constant, then the regression coefficients will also change.

Difference between Correlation and Regression

Basis for Comparison	Correlation	Regression
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).

Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

General Difference between Correlation and Regression

Sl.No.	Correlation	Regression
1	Correlation is the relationship between variables. It is expressed numerically	Regression means going back. The average relationship between variables is given as an equation
2	Between two variables, none is identified as independent or dependent variable.	One of the variable is independent and other is dependent variable in any particular context
3	Correlation does not mean causation. One variable need not be the cause and the other effect	Independent variable may be 'the cause and depndent variable,' the effect'.
4	There is spurious or non sense correlation	Regression is considered. Regression is considered only when the variables are related.
5	Correlation co efficient is independent change of origin and scale.	Correlation co efficient is independent change of origin but are affected by change of scale.
6	Correlation coefficient is a number between -1 to +1	The two regression coefficients have the same sign, + or-. One of them can be greater than unity. But they can not be greater than one numerically simultaneously
7	Correlation coefficient is not in any unit of measurement	The regression coefficients is in the unit of measurement in the dependent variable
8	Correlation coefficient indicates the direction of co variation and the closeness of the linear relation between two variables	Regression equations give the value of dependent variable corresponding to any value of the independent variable.

Taking the new pairs of values, correlation can be calculated in the same manner as discussed carlier.

Illustration 27. The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after 2 months. Allowing for this time lag, calculate coefficient of correlation.

Months	*Advertising Expenditure*	*Sales*	*Months*	*Advertising Expenditure*	*Sales*
Jan.	50	1,200	July	140	2,400
Feb.	60	1,500	Aug.	160	2,600
March	70	1,600	Sep.	170	2,800
April	90	2,000	Oct.	190	2,900
May	120	2,200	Nov.	200	3,100
June	150	2,500	Dec.	250	3,900

Solution. Allow for a time lag of 2 months, *i.e.*, link advertising expenditure of January with sales for March, and so on.

CALCULATION OF CORRELATION COEFFICIENT

Months	*Advertising Expenditure* X	$\frac{(X-\overline{X})}{10}$ x	x^2	*Sales* Y	$\frac{(Y-\overline{Y})}{100}$ y	y^2	xy
Jan.	50	−7	49	1,600	−10	100	70
Feb.	60	−6	36	2,000	−6	36	36
March	70	−5	25	2,200	−4	16	20
April	90	−3	9	2,500	−1	1	3
May	120	0	0	2,400	−2	4	0
June	150	+3	9	2,600	0	0	0
July	140	+2	4	2,800	+2	4	4
Aug.	160	+4	16	2,900	+3	9	12
Sep.	170	+5	25	3,100	+5	25	25
Oct.	190	+7	49	3,900	+13	169	91
	$\Sigma X =$ 1,200	$\Sigma x = 0$	$\Sigma x^2 =$ 222	$\Sigma Y =$ 26,000	$\Sigma y = 0$	$\Sigma y^2 =$ 364	$\Sigma xy =$ 261

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

$$\overline{X} = \frac{1{,}200}{10} = 120,$$

$$\overline{Y} = \frac{26{,}000}{10} = 2{,}600$$

$$\Sigma xy = 261, \ \Sigma x^2 = 222, \ \Sigma y^2 = 364$$

$$r = \frac{261}{\sqrt{222 \times 364}} = \frac{261}{284{\cdot}267}$$

$$= +\,0{\cdot}918.$$

Illustration 28. Find the coefficient of the correlation for the following :

Cost :	39	65	62	90	82	75	25	98	36	78
Sales :	47	53	58	86	62	68	60	91	51	84

(*B. Com., Madras Univ., 1997*)

Solution. CALCULATION OF COEFFICIENT OF CORRELATION BY KARL PEARSON'S METHOD

X	$(X-\bar{X})$	x^2	Y	$(Y-\bar{Y})$	y^2	xy
	$\bar{X}=65$			$\bar{Y}=66$		
	x			y		
39	−26	676	47	−19	361	+494
65	0	0	53	−13	169	0
62	−3	9	58	−8	64	+24
90	+25	625	86	+20	400	+500
82	+17	289	62	−4	16	−68
75	+10	100	68	+2	4	+20
25	−40	1600	60	−6	36	+240
98	+33	1089	91	+25	625	+825
36	−29	841	51	−15	225	+435
78	+13	169	84	+18	324	+234
$\Sigma X = 650$	$\Sigma x = 0$	$\Sigma x^2 = 5398$	$\Sigma Y = 660$	$\Sigma y = 0$	$\Sigma y^2 = 2224$	$\Sigma xy = 2704$

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

$$= \frac{2704}{\sqrt{5398 \times 2224}}$$

$$= \frac{2704}{3464 \cdot 85} = + 0 \cdot 78.$$

Illustration 29. The following data relate to age of employees and the number of days they were reported sick in a month.

	Employees									
	1	2	3	4	5	6	7	8	9	10
Age (X)	30	32	35	40	48	50	52	55	57	61
Sick days (Y)	1	0	2	5	2	4	6	5	7	8

Calculate Karl Pearson's Coefficient of Correlation and interpret it.

(*B. Com., Kashmir Univ.,* 1997)

Solution. CALCULATION OF KARL PEARSON'S COEFFICIENT OF CORRELATION

Age	$(X-\bar{X})$		*Sick days*	$(Y-\bar{Y})$		
X	x	x^2	Y	y	y^2	xy
30	−16	256	1	−3	9	+48
32	−14	196	0	−4	16	+56
35	−11	121	2	−2	4	+22
40	−6	36	5	+1	1	−6
48	+2	4	2	−2	4	−4
50	+4	16	4	0	0	0
52	+6	36	6	+2	4	+12
55	+9	81	5	+1	1	+9
57	+11	121	7	+3	9	+33
61	+15	225	8	+4	16	+60
$\Sigma X = 460$	$\Sigma x = 0$	$\Sigma x^2 = 1092$	$\Sigma Y = 40$	$\Sigma y = 0$	$\Sigma y^2 = 64$	$\Sigma xy = 230$

$$\overline{X} = \frac{460}{10} = 46, \quad \overline{Y} = \frac{40}{10} = 4$$

Since the actual means of X and Y are not in fraction, we can use the direct method of calculating correiation coefficient.

$$r = \frac{\Sigma\, x\, y}{\sqrt{\Sigma\, x^2 \times \Sigma\, y^2}} = \frac{230}{\sqrt{1092 \times 64}}$$

$$= \frac{230}{264{\cdot}363} = +\,0{\cdot}87.$$

There is a high degree of positive correlation between age and number of days reported sick.

Illustration 30. Calculate Karl Pearson's Coefficient of Correlation between age and playing habits from the data given below. Also calculate probable error and comment on the value :

Age :	20	21	22	23	24	25
No. of Students :	500	400	300	240	200	160
Regular Players :	400	300	180	96	60	24

(MBA, HPU, 1998; MBA, Kumaun Univ., 2001)

Solution. Let us first find the percentage of regular players and then calculate correlation between age and percentage.

Age *X*	*(X − 22)* d_x	d_x^2	*No. of Students*	*Regular Players*	*% of Regular Players* *Y*	*(Y − 50)* d_y	d_y^2	$d_x\,d_y$
20	−2	4	500	400	80	+30	900	−60
21	−1	1	400	300	75	+25	625	−25
22	0	0	300	180	60	+10	100	0
23	+1	1	240	96	40	−10	100	−10
24	+2	4	200	60	30	−20	400	−40
25	+3	9	160	24	15	−35	1225	−105
$\Sigma X =$ 135	$\Sigma\, d_x = 3$	$\Sigma\, d_x^2 = 19$			$\Sigma\, Y = 300$	$\Sigma\, d_y = 0$	$\Sigma\, d_y^2 =$ 3350	$\Sigma\, d_x\, d_y$ $= -\,240$

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

$$= \frac{(6 \times -\,240) - (3 \times 0)}{\sqrt{6 \times 19 - (3)^2}\ \sqrt{6 \times 3350}}$$

$$= \frac{-\,1440}{\sqrt{105 \times 20100}} = \frac{-\,1440}{1452{\cdot}756} = -\,0{\cdot}991$$

$$\text{P.E.}_{\,r} = 0{\cdot}6745\ \frac{1 - r^2}{\sqrt{N}}$$

$$= 0{\cdot}6745\ \frac{1 - ({\cdot}991)^2}{\sqrt{6}}$$

$$= \frac{0{\cdot}6745 \times {\cdot}018}{2{\cdot}449} = 0{\cdot}005.$$

Illustration 31. The ranks of the same 15 students in two subjects A and B are given below. The two numbers within brackets denote the ranks of the same student in A and B respectively.

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

Find the Spearman's Rank Correlation Coefficient

(MBA, Sukhadia Univ., 1998)

Solution. CALCULATION OF SPEARMAN'S RANK CORRELATION COEFFICIENT

R_A	R_B	$(R_A - R_B)^2$ D^2
1	10	81
2	7	25
3	2	1
4	6	4
5	4	1
6	8	4
7	3	16
8	1	81
9	11	4
10	15	25
11	9	4
12	5	49
13	14	1
14	12	4
15	13	4
		$\Sigma D^2 = 304$

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

$$= 1 - \frac{6 \times 304}{15^3 - 15} = 1 - \frac{1824}{3360} = 1 - 0{\cdot}543 = 0{\cdot}457.$$

Illustration 32. With the following data in 6 cities calculate the coefficient of correlation by Pearson's method between the density of population and death rate :

City	*Area in kilometres*	*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

(B.Com., Sukhadia Univ., 1998)

Solution. First we will calculate density of population and death rate and denote them by *X* and *Y*.

$$\text{Density} = \frac{\text{Population}}{\text{Area}}; \text{ Death rate} = \frac{\text{No. of Deaths}}{\text{Population}} \times 1000$$

City	*Density* X	$(X-450)/100$ x	x^2	*Death rate* Y	$(Y-15)$ y	y^2	xy
A	200	–2.5	6.25	10	–5	25	+12.5
B	500	+0.5	0.25	16	+1	1	+0.5
C	400	–0.5	0.25	14	–1	1	+0.5
D	700	+2.5	6.25	20	+5	25	+12.5
E	600	+1.5	2.25	17	+2	4	+3.0
F	300	–1.5	2.25	13	–2	4	+3.0
		$\Sigma x = 0$	$\Sigma x^2 = 17.5$	$\Sigma Y = 90$	$\Sigma y = 0$	$\Sigma y^2 = 60$	$\Sigma xy = 32$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}} = \frac{32}{\sqrt{17{\cdot}5 \times 60}} = \frac{32}{32{\cdot}404} = +\,0{\cdot}988$$

The value of this coefficient interpreted in the same way as Karl Pearson's correlation coefficient, ranges between +1 and −1. When r_2 is +1 there is complete agreement in the order of the ranks and the ranks are in the same direction. When r_2 is −1 there is complete agreement in the order of the ranks and they are in opposite directions. This shall be clear from the following :

R_1	R_2	D $(R_1—R_2)$	D^2	R_1	R_2	D $(R_1—R_2)$	D^2
1	1	0	0	1	3	−2	4
2	2	0	0	2	2	0	0
3	3	0	0	3	1	2	4
			$\Sigma D^2=0$				$\Sigma D^2=8$

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

$$= 1 - \frac{6 \times 0}{3^3 - 2} = 1 - 0 = 1 \qquad\qquad = 1 - \frac{6 \times 8}{3^3 - 3} = 1 - 2 = -1$$

Features of Spearman's Correlation Coefficient

- The sum of the differences of ranks between two variables shall be zero. Symbolically, $\Sigma\, d = 0$
- Spearman's correlation coefficient is distribution-free or non-parametric because no strict assumptions are made about the form of population from which sample observations are drawn.
- The Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks. Hence, it can be interpreted in the same manner as Pearsonian correlation coefficient.

In rank correlation we may have two types of problems :

- Where ranks are given.
- Where ranks are not given.

Where Ranks are Given Where actual ranks are given to us the steps requird for computing rank correlation are :

(*i*) Take the differences of the two ranks, *i.e.*, $(R_1 - R_3)$ and denote these differences by D.

(*ii*) Square these differences and obtain the total $\Sigma\, D^2$.

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

Illustration 16. The ranking of 10 students in two subjects A and B are as follows :

A	*B*	*A*	*B*
6	3	4	6
5	8	9	10
3	4	7	7
10	9	8	5
2	1	1	2

Calculate rank correlation coefficient.

(*B. Com., JU.* 1997)

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

R_1	R_2	$(R_1 - R_2)^2$ D^2
6	3	9
5	8	9
3	4	1
10	9	1
2	1	1
4	6	4
9	10	1
7	7	0
8	5	9
1	2	1
		$\Sigma D^2 = 36$

$$R = 1 - \frac{6\,\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 36}{10^3 - 10}$$

$$= 1 - \frac{216}{990} = 0{\cdot}782$$

Illustration 17. Two ladies were asked to rank 7 different types of lipsticks. The ranks given by them are as follows :

Lipsticks	*A*	*B*	*C*	*D*	*E*	*F*	*G*
Neelu	2	1	4	3	5	7	6
Neena	1	3	2	4	5	6	7

Calculate Spearman's rank correlation coefficient.

Solution. CALCULATION OF SPEARMAN'S RANK CORRELATION COEFFICIENT

X R_1	Y R_2	$(R_1 - R_2)$ D	D^2
2	1	+1	1
1	3	−2	4
4	2	+2	4
3	4	−1	1
5	5	0	0
7	6	+1	1
6	7	−1	1
			$\Sigma D^2 = 12$

$$R = 1 - \frac{6\,\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 12}{7^3 - 7} = 1 - 0{\cdot}214 = 0{\cdot}786.$$

Illustration 18. Ten competitors in a beauty contest are ranked by three judges in the following order :

Ist 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

Year	*Debenture price*	*Share price*
1	97.8	73.2
2	99.2	85.8
3	98.8	78.9
4	98.3	75.8
5	98.4	77.2
6	96.7	87.2
7	97.1	83.8

Using rank correlation method, determine the relationship between debenture prices and share prices. *(B.Com., Calicut Univ., 1997)*

Solution. CALCULATION OF RANK CORRELATION COEFFICIENT

X	*Rx*	*Y*	*Ry*	$(R_x — R_y)^2$ D^2
97.8	3	73.2	1	4
99.2	7	85.8	6	1
98.8	6	78.9	4	4
98.3	4	75.8	2	4
98.4	5	77.2	3	4
96.7	1	87.2	7	36
97.1	2	83.8	5	9
				$\Sigma D^2 = 62$

$$R = 1 - \frac{6\,\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 62}{7^3 - 7} = 1 - \frac{372}{336} = 1 - 1{\cdot}107 = -\,0{\cdot}107.$$

Illustration 20. Calculate Spearman's coefficient of correlation between marks assigned to ten students by judges *X* and *Y* in a certain competitive test as shown below :

S. No. :	1	2	3	4	5	6	7	8	9	10
Marks by judge X :	52	53	42	60	45	41	37	38	25	27
Marks by judge Y :	65	68	43	38	77	48	35	30	25	50

Solution. First assign ranks and then calculate rank correlation coefficient

COMPUTATION OF SPEARMAN'S COEFFICIENT OF CORRELATION

Marks by judge X	*Rx*	*Marks by judge Y*	*Ry*	$(R_x - R_y)$ D^2
52	8	65	8	0
53	9	68	9	0
42	6	43	5	1
60	10	38	4	36
45	7	77	10	9
41	5	48	6	1
37	3	30	3	0
38	4	32	2	4
25	1	25	1	0
27	2	50	7	25
				$\Sigma D^2 = 76$

$$R = 1 - \frac{6\,\Sigma D^2}{N^3 - N} = 1 - \frac{6 \times 76}{10^3 - 10} = 1 - 0{\cdot}461 = -\,0{\cdot}539.$$

Illustration 21. Calculate the coefficient of correlation from the following data by the Spearman's Rank difference method :

Price of Tea (Rs.)	*Price of Coffee (Rs.)*	*Price of Tea (Rs.)*	*Price of Coffee (Rs.)*
75	120	60	110
88	134	80	140
95	150	81	142
70	115	50	100

Solution. CALCULATION OF SPEARMAN'S CORRELATION COEFFICIENT

Price of Tea (Rs.)	R_1	*Price of Coffee (Rs.)*	R_2	$(R_1 - R_2)^2$ D^2
75	4	120	4	0
88	7	134	5	4
95	8	150	8	0
70	3	115	3	0
60	2	110	2	0
80	5	140	6	1
81	6	142	7	1
50	1	100	1	0
				$\Sigma D^2 = 6$

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

$$= 1 - \frac{36}{512 - 8} = 1 - 0{\cdot}071 = +\,0{\cdot}929$$

Equal Ranks In some cases it may be found necessary to rank two or more individuals or entries as equal. In such a case it is customary to give each individual an average rank. Thus, if two individuals are ranked equal at fifth place, they are each given the rank $\frac{5+6}{2}$, that is 5.5 while, if three are ranked equal at fifth place, they are given the rank $\frac{5+6+7}{3} = 6$. In other words, where two or more items are to be ranked equal, the rank assigned for purposes of calculating coefficient of correlation is the average of the ranks which these individuals would have got had they differed slightly from each other.

Where equal ranks are assigned to some entries an adjustment in the above formula for calculating the rank coefficient of correlation is made.

The adjustment consists of adding $\frac{1}{12}(m^3 - m)$ to the value of ΣD^2, where M stands for the number of items whose ranks are common. If there are more than one such group of items with common rank, this value is added as many times the number of such groups. The formula can thus be written

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

Illustration 22. Obtain the rank correlation coefficient between the variables X and Y from the following pairs of observed values.

X:	50	55	65	50	55	60	50	65	70	75
Y:	110	110	115	125	140	115	130	120	115	160

(B.Com. Mangalore Univ., 1997)

Solution. For finding ranks correlation coefficient first rank two various values. Taking lowest as 1 and next higher as 2, etc..

X	*Rank X* R_1	Y	*Rank Y* R_2	$(R_1 - R_2)^2$ D^2
50	2	110	1.5	0.25
55	4.5	110	1.5	9.00
65	7.5	115	4	12.25
50	2	125	7	25.00
55	4.5	140	9	20.25
60	6	115	4	4.00
50	2	130	8	36.00
65	7.5	120	6	2.25
70	9	115	4	25.00
75	10	160	10	00.00
				$\Sigma D^2 = 134$

It may be noted that in series X, 50 has repeated thrice ($m = 3$), 55 has been repeated twice ($m = 2$), 65 has been repeated twice ($m = 2$). In series Y, 110 has been repeated twice ($m = 2$) and 115 thrice ($m = 3$)

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

$$R = 1 - \frac{6\left\{134 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(3^3 - 3)\right\}}{10^3 - 10}$$

$$= 1 - \frac{6\,[134 + 2 + \cdot 5 + \cdot 5 + \cdot 5 + 2]}{990}$$

$$= 1 - \frac{6\,(139\cdot 5)}{990} = 1 - \frac{837}{990} = 1 - \cdot 845 = 0\cdot 155$$

erits and Limitations of the Rank Method

erits. The merits of the Rank Method can be discussed here :

- This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answers obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, *i.e.*, all the items are different.

It is clear from this example that answer would come out to be the same whether we take deviations from actual means or assumed means.

Graphing Regression Lines It is quite easy to graph the regression lines once they have been computed. All one has to do is to—

(a) choose any two values (preferably well apart) for the unknown variable on the right-hand side of the equation,

(b) compute the other variable,

(c) plot the two pairs of values, and

(d) draw a straight line through the plotted points.

Illustration 4. Show graphically the regression equations of illustration 3.

Solution. (a) *Regression Line of Y on X* [$Y = 11{\cdot}9 - 0{\cdot}63\,X$].

(i) *Let* $X = 2$, $Y = 11{\cdot}9 - 0{\cdot}65\ (2) = 11{\cdot}9 - 1{\cdot}3 = 10{\cdot}6$.

(ii) *Let* $X = 10$, $Y = 11{\cdot}9 - 0{\cdot}65 \times 10 = 5{\cdot}4$.

These points and the regression line through them are shown on the graph on the next page.

(b) *Regression line of X on Y* ($X = 16{\cdot}4 - 1{\cdot}3\ Y$)

(i) *Let* $Y = 10$

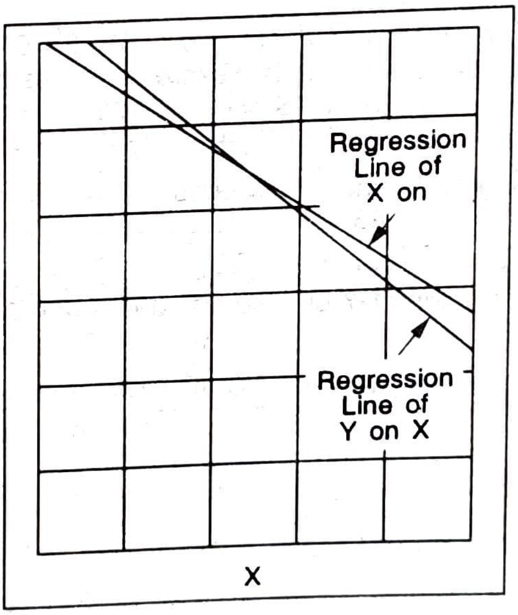
$\therefore$ $x = 16{\cdot}4 - 1{\cdot}3\ (10) = 16{\cdot}4 - 13 = 3{\cdot}4$

(ii) *Let* $Y = 6$

$\therefore$ $x = 16{\cdot}4 - 1{\cdot}3\ (6) = 16{\cdot}4 - 7{\cdot}8 = 8{\cdot}6$

Illustration 5. From the data of illustration 1, obtain regression equations taking deviations from 5 in case of X and 7 in case of Y :

These points and the regression line through them are shown in the graph below :



Thus the value of regression coefficient comes out to be the same.

Illustration 6. The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales in thousand rupees :

Salesmen Intelligence :	A	B	C	D	E	F	G	H	I
Test Scores :	50	60	50	60	80	50	80	40	70
Weekly Sales :	30	60	40	50	60	30	70	50	60

(a) Obtain the regression equation of sales on intelligence test scores of the salsemen

(b) If the intelligence test score of a salesman in 65, what would be his expected weekly sales?

(*M. Com., HPU.,* 1996)

Solution. Let intelligence test score be denoted by X and weekly sales by Y.

CALCULATION OF REGRESSION EQUATIONS

X	$(X-60)$ x	x^2	Y	$(Y-50)$ y	y^2	xy
50	−10	100	30	−20	400	+200
60	0	0	60	+10	100	0
50	−10	100	40	−10	100	+100
60	0	0	50	0	0	0
80	+20	400	60	+10	100	+200
50	−10	100	30	−20	400	+200
80	+20	400	70	+20	400	+400
40	−20	400	50	0	0	0
70	+10	100	60	+10	100	+100
$\Sigma X=540$	$\Sigma x=0$	$\Sigma x^2=1600$	$\Sigma Y=450$	$\Sigma y=0$	$\Sigma y^2=1600$	$\Sigma xy=1200$

Regression equation of Y on X : $Y-\overline{Y}=r\frac{\sigma_y}{\sigma_x}(X-\overline{X})$

$$r\frac{\sigma_y}{\sigma_x}=\frac{\Sigma xy}{\Sigma x^2}=\frac{1200}{1600}=0{\cdot}75$$

$$\overline{X}=\frac{\Sigma X}{N}=\frac{540}{9}=60,\ \overline{Y}=\frac{\Sigma Y}{N}=\frac{450}{9}=50$$

$$Y-50=0{\cdot}75\,(X-60)$$

$$Y-50=0{\cdot}75\,X-45 \quad \text{or} \quad Y=5+0{\cdot}75\,X$$

Expected weekly sales when intelligence test score of a salesman is 65

$$Y=5+0{\cdot}75\,X. \qquad \text{Putting } X=65$$

$$Y=0{\cdot}75\times(65)+5=48{\cdot}75+5=53{\cdot}75$$

Illustration 7. The following table shows the ages (X) and blood pressure (Y) of 8 persons.

X :	52	63	45	36	72	65	47	25
Y :	62	53	51	25	79	43	60	33

Obtain the regression equation of Y on X and find the expected blood pressure of a person who is 49 years old.

(*B. Com., Bombay Univ.*, 1996)

Solution. CALCULATION OF REGRESSION EQUATION OF Y ON X

X	$(X-50)$ d_x	d_x^2	Y	$(Y-50)$ d_y	d_y^2	d_xd_y
52	+2	4	62	+12	144	+24
63	+13	169	53	+3	9	+39
45	−5	25	51	+1	1	−5
36	−14	196	25	−25	625	+350
72	+22	484	79	+29	841	+638
65	+15	225	43	−7	49	−105
47	−3	9	60	+10	100	−30
25	−25	625	33	−17	289	+425
$\Sigma X=405$	$\Sigma d_x=5$	$\Sigma d_x^2=1737$	$\Sigma Y=406$	$\Sigma d_y=6$	$\Sigma d_y^2=2058$	$\Sigma d_xd_y=1336$

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

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{406}{8} = 50{\cdot}75\ ;\ \overline{X} = \frac{\Sigma X}{N} = \frac{405}{8} = 50{\cdot}625$$

$$r\frac{\sigma_y}{\sigma_x} = \frac{N\Sigma d_x d_y - \Sigma d_x \Sigma d_y}{N\Sigma d_x^2 - (\Sigma d_x)^2} = \frac{(8)(1336) - (5)(6)}{(8)(1737) - (5)^2} = \frac{10688 - 30}{13896 - 25} = 0{\cdot}768$$

$$Y - 50{\cdot}75 = 0{\cdot}768\,(X - 50{\cdot}625)$$

$$Y - 50{\cdot}75 = 0{\cdot}768\,X - 38{\cdot}88 \quad \text{or} \quad Y = 11{\cdot}87 + 0{\cdot}768\,X$$

$$Y_{49} = 11{\cdot}87 + 0{\cdot}768\,(49) = 49{\cdot}502$$

Thus, the expected blood pressure of a person who is 49 years old shall be 49·5.

Illustration 8. In a correlation study the following values are obtained :

	X	Y
Mean	65	67
Standard Deviation	2·5	3·5
Coefficient of Correlation	0·8	

Find the two regression equations that are associated with the above values.

(B. com., Kashmir Univ., 1996; MBA. HPU, 2000)

Solution. The two regression equations are :

Regression Equation of X on Y : $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

$$\overline{X} = 65,\ r = 0{\cdot}8,\ \sigma_x = 2{\cdot}5,\ \sigma_y = 3{\cdot}5,\ \overline{Y} = 67$$

Substituting the values : $X - 65 = 8\,\frac{2{\cdot}5}{3{\cdot}5}\,(Y - 67)$

$$X - 65 = {\cdot}5714\,(Y - 67)$$

$$X - 65 = {\cdot}5714\,Y - 38{\cdot}28 \quad \text{or} \quad X = 26{\cdot}72 + 0{\cdot}5714\,Y$$

Regression Equation of Y on X : $Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$

$$Y - 67 = 8\,\frac{3{\cdot}5}{2{\cdot}5}\,(X - 65)$$

$$Y - 67 = 1{\cdot}12\,(X - 65)$$

$$Y - 67 = 1{\cdot}12\,X - 72{\cdot}8 \quad \text{or} \quad Y = -\,5{\cdot}8 + 1{\cdot}12\,X.$$

Illustration 9. In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :

Variance of $X = 9$

Regression equations $8X - 10Y + 66 = 0$

$40X - 18Y = 214$

Find on the basis of the above information

(*i*) The mean values of X and Y,

(*ii*) Coefficient of correlation between X and Y, and

(*iii*) Standard deviation of Y.

(B. Sc., Madurai-Kamaraj Univ., ; M. Com., M.D. Univ.,; ***M.Com., Bhopal, 1999; M.Com., Allahabad Univ., 2001)***

Solution.

(*i*) *The Mean values of X and Y* : $8X - 10Y = -66$ ...(*i*)

$40X - 18Y = 214$ ...(*ii*)

Multiplying equation (*i*) by 5 $\quad 40X - 50Y = -330$

$$\begin{array}{r} 40X - 18Y = 214 \\ -\quad\ \ +\quad\ \ - \\ \hline -32Y = -544 \end{array}$$

$$Y = 17 \text{ or } \overline{Y} = 17$$

Substituting the value of Y in eq. (*i*) : $8X - 10 \times 17 = -66$

$$8X = -66 + 170$$

$$8X = 104 \qquad \therefore\ X = 13 \text{ or } \overline{X} = 13$$

(*ii*) For finding out the correlation coefficient, we will have to find out the regression coefficient. Since we do not know which of the two regression equations is the equation of X on Y, we make an assumption. Let us take eq. (*i*) as the regression equation of X on Y.

$$8\,X = -\,66 + 10\,Y$$

$$X = -\frac{66}{8} + \frac{10}{8}\,Y; \quad \text{or} \quad b_{xy} = \frac{10}{8} = 1{\cdot}25$$

From eq. (*ii*) we can calculate b_{yx} $40\,X - 18\,Y = 214$

or, $-\,18\,Y = 214 - 40\,X$

$$Y = -\frac{214}{18} + \frac{40}{18}\,X \quad \text{or} \quad b_{yx} = \frac{40}{18}$$

Since both the regression coefficients are exceeding 1, our assumption is wrong. Hence, the first equation is equation of *Y* on *X*.

From eq. (*i*)

$$-10Y = -8X - 66$$

$$Y = -\frac{8}{10}\,X + 6{\cdot}6 \quad \text{or} \quad b_{yx} = \frac{8}{10}$$

From eq. (*ii*) $b_{xy} = \frac{18}{40} = 0{\cdot}45$

$$r = \sqrt{\frac{8}{10} \times \frac{18}{40}} = \sqrt{0{\cdot}36} = 0{\cdot}6$$

$$\sigma_x = \sqrt{9} = 3\,;\ b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

$$0{\cdot}45 = 0{\cdot}6\,\frac{3}{\sigma_y} \quad \text{or} \quad 0{\cdot}45\,\sigma_y = 1{\cdot}8 \quad \text{or} \quad \sigma_y = \frac{1{\cdot}8}{{\cdot}45} = 4$$

Hence, standard deviation of *Y* is 4.

Illustration 10. For 50 students of a class the regression equation of marks in Statistics (*X*) on the marks in Accountancy (*Y*) is $3\,Y - 5\,X + 180 = 0$. The mean marks in Accountancy is 44 and variance of marks in Statistics is 9/16th of the variance of marks in Accountancy. Find the mean marks in Statistics and the coefficient of correlation between marks in the two subjects. *[M. Com., Madurai,* 1993 ; *B. Com.* (*H*), *Delhi Univ.*, 1994]

Solution. We are given

$$3\,Y - 5\,X + 180 = 0 \quad \text{or} \quad 3\,Y + 180 = 5\,X$$

X represents marks in Statistics and *Y* marks in Accountancy. When $Y = 44$, *X* will be given by

$$5\,X = (3)\,(44) + 180 = 0\,;\ 5\,X = 132 + 180 \text{ or } X = \frac{312}{5} = 62{\cdot}4$$

Hence, the mean marks in Statistics are 62·4.

For calculating coefficient of correlation we know that

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

Regression coefficient of X on Y from the given equation is

$$5\,X = 3\,Y + 180 \quad \text{or} \quad X = 0{\cdot}6\,Y + 36$$

$$\therefore \quad b_{xy} = 0{\cdot}6\,;\ r\frac{\sigma_x}{\sigma_y} = \frac{\sqrt{9}}{\sqrt{16}} \quad \text{given}$$

$$\therefore \quad 0{\cdot}6 = r\frac{\sqrt{9}}{\sqrt{16}} \quad \text{or} \quad 0{\cdot}6 = r\frac{3}{4}$$

Hence, $3r = 2{\cdot}4 \quad \therefore \quad r = +\,0{\cdot}8.$

Illustration 11. You are given the following data :

	X	*Y*
Arithmetic mean	36	85
Standard Deviation	11	8

Correlation coefficient between *X* and *Y* = 0·66

(*i*) Find the two Regression Equations, and

(*ii*) Estimate the value of *X* when *Y* = 75.

([*B. A. (Hons.). Econ. Delhi Univ. B. Com Guwahati Univ..* 1998] lution.

(i) *Regression Equation of X on Y* : $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

$$\overline{X} = 36,\ r = 0{\cdot}66, \sigma_x = 11,\ \sigma_y = 8,\ \overline{Y} = 85$$

$$X - 36 = 0{\cdot}66\ \frac{11}{8}\ (Y - 85)$$

$$X - 36 = {\cdot}9075\ (Y - 85)$$

$$X = {\cdot}9075\ Y - 77{\cdot}1375 + 36 \quad \text{or } X = -\ 41{\cdot}1375 + {\cdot}9075\ Y$$

egression Equation of Y on X : $Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$

$$Y - 85 = {\cdot}66\ \frac{8}{11}\ (X - 36)$$

$$Y - 85 = {\cdot}48\ (X - 36)$$

$$\underline{Y - 85 = {\cdot}48\ X - 17{\cdot}28 \text{ or }} \quad Y = 67{\cdot}72 + 0{\cdot}48\ X$$

) From the regression equation of X on Y, we can find out the estimated value of X when 75 ; $X = {\cdot}9075\ (75) - 41{\cdot}1375$

$$= 68{\cdot}0625 - 41{\cdot}1375 = 26{\cdot}925 \quad \text{or} \quad Y_{75} = 26{\cdot}925.$$

stration 12. For certain X and Y series which are correlated, the two lines of regres- are :

$$5\ X - 6\ Y + 90 = 0$$

$$15\ X - 8\ Y - 130 = 0$$

nd the means of the two series and the correlation coefficient. *(M. Com., M.D. Univ., 1998)*

tion : (i) *Finding mean of the two series* :

$$5\ X - 6\ Y = -\ 90 \qquad ...(i)$$

$$15\ X - 8\ Y = 130 \qquad ...(ii)$$

ltiplying eq. (i) by 3, $15\ X - 18\ Y = -\ 270$

$$15\ X - 8\ Y = 130$$

$$-\qquad +\qquad -$$

$$-\ 10\ Y = -\ 400$$

$$Y = 40 \quad \text{or} \quad \overline{Y} = 40$$

tting the value of Y in eq. (i), $5\ X - 6\ (40) = -\ 90$

$$5\ X = -\ 90 + 240$$

$$5\ X = 150\text{r } X = 30 \text{ or } \overline{X} = 30$$

Finding correlation coefficient. Let us assume that eq. (i) is the regression equation of X on

$$5\ X = 6\ Y - 90$$

$$X = \frac{6}{5}\ Y - 18 \text{ or } b_{xy} = \frac{6}{5}$$

king eq. (ii) as the eq. of Y on X, $-\ 8\ Y = -\ 15\ X + 130$

$$8\ Y = 15\ X - 130$$

$$Y = \frac{15}{8}\ X - \frac{130}{8} \quad \text{or } b_{yx} = \frac{15}{8}$$

ce both the regression coefficients are exceeding one, our assumption is wrong. nce, eq. (i) is the regression eq. of Y on X

$$-\ 6\ Y = -\ 5\ X - 90 \quad \text{or } 6\ Y = 5\ X + 90$$

$$Y = \frac{5}{6}\ X + 15 \quad \text{or } b_{yx} = \frac{5}{6}$$

(ii) is the regression eq. of X on Y ; $15\ X = 130 + 8\ Y$

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

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{\frac{8}{15} \times \frac{5}{6}} = 0{\cdot}667$$