CORRELATION

Introduction

Correlation means a relation between two groups. In statistics, it is the measure to indicate the relationship between two variables in which, with changes in the values of one variable, the values of another variable also change.

These variables may be related to one item or may not be related to one item but have dependence on the other due to some reason. For example, the data on height and weights of a group of people would relate to each member of the group but prices of sugar and sugarcane are two different series altogether but there would be some relation between the values of the two, prices of sugar depending upon the prices of sugarcane.

Correlation provides a tool into the hands of decision-makers because it provides better understanding of the trends and their dependence on other factors so that the range of uncertainties associated with decision-making is reduced.

Definition

The term correlation indicates the relationship between two variables in which with changes in the value of one variable, the values of the other variable also change. Correlation has been defined by various eminent statisticians, mathematicians, and economists. Some of the important definitions of correlation are given below:

(1) According to La Yun Chow, “Correlation analysis attempts to determine the degree of relationship between variables.”

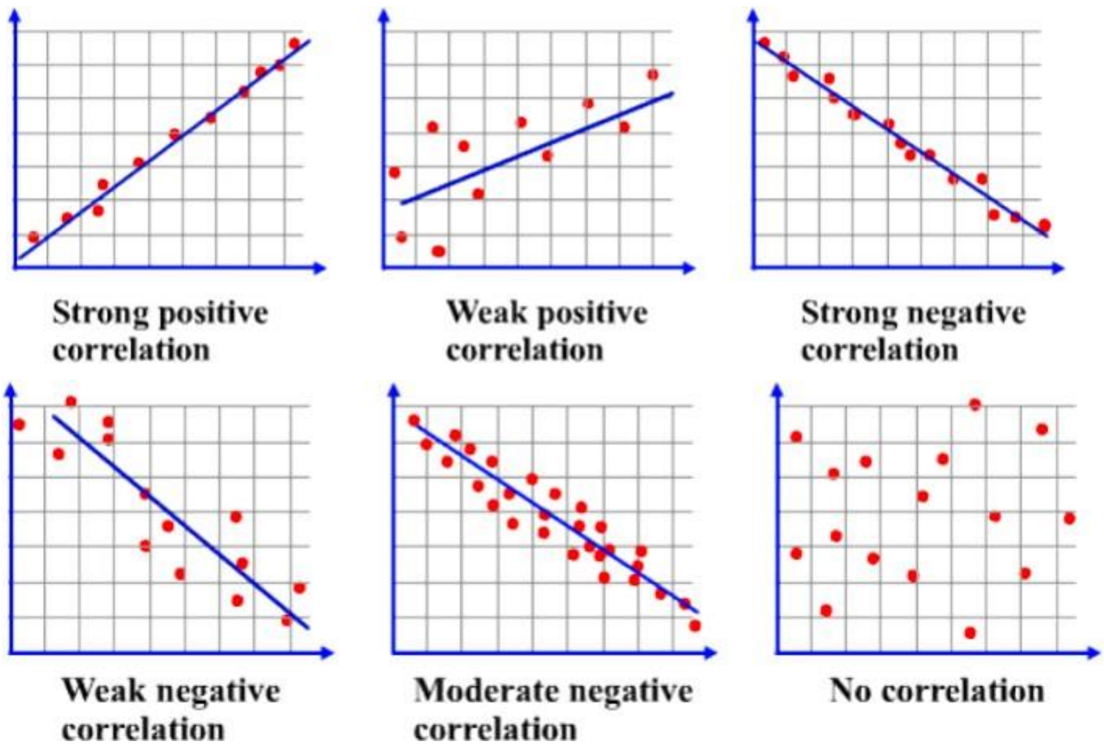
(2) Croxton and Lowden define correlation as, “When the relationship of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as correlation.”

(3) As per Prof. Boddington, “Whenever some definite connection exists between two or more groups, classes or series of data, there is said to be correlation.”

Scatter Diagram

This is the simplest method of studying the correlation between two variables. The two variables x and y are taken on the X and Y axes of a graph paper. Each pair of x and y value we mark a dot and we get as many points as the number of pairs of observation. By looking through the scatter of points, we can form an idea as whether the variables are related or not.

- If all the plotted *points lie on a straight line* rising from the lower left-hand corner to the upper right-hand corner, correlation is said to be ***perfectly positive***.
- If all the plotted points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner of the diagram, correlation is said to be ***perfectly negative***.
- If all the plotted points fall in a narrow line and the points are rising from the lower left hand corner to the upper right hand corner of the diagram, there is degree of ***positive correlation*** between variables.
- If the plotted points fall in a narrow bank and the points are lying from the upper left-hand corner to the right-hand corner, there high degree of ***negative correlation***.
- If the plotted points lie scattered all over the diagram, there is ***no correlation*** between the two variables.



Correlation and Causation

Correlation is a statistical technique which tells us how strongly the pair of variables are linearly related and change together. It does not tell us why and how behind the relationship but it just says the relationship exists.

Example: Correlation between ice cream sales and sunglasses sold.

Causation takes a step further than correlation. It says any change in the value of one variable will cause a change in the value of another variable, which means one variable makes other to happen. It is also referred as cause and effect.

Example: When a person is exercising then the amount of calories burning goes up every minute. The former is causing the latter to happen.

Types of Correlation

Correlation can be classified as given ahead:

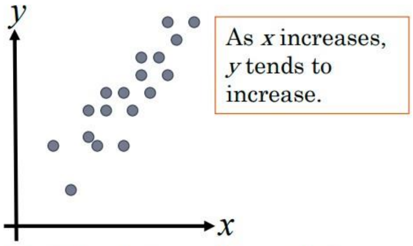
I. Positive, Negative and Zero Correlation:
II. Simple, Partial and Multiple Correlation:
III. Linear and Non-linear Correlation:

I. Positive, Negative and Zero Correlation

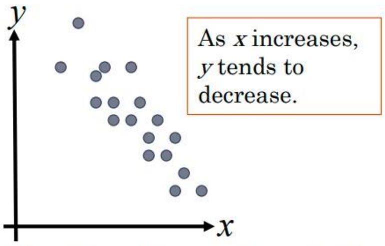
Whether correlation is positive or negative would depend upon the direction of change of the variable.

Positive Correlation:

When the values of the two variables move in the same direction, i.e., an increase in one is associated with an increase in other, or vice versa, the correlation is said to be positive. For example, the correlation between heights and weights of a group of persons is a positive correlation.



Positive Linear Correlation



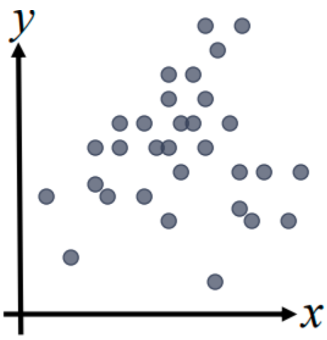
Negative Linear Correlation

Negative Correlation:

If the values of two variables move in the opposite directions i.e., an increase in the value of one variable is associated with fall in other, or vice versa, the correlation is said to be negative. For example, the correlation between the temperature and pressure is a negative correlation.

Zero Correlation:

When we don't find any relationship between the variables then, it is said to be zero correlation. It means a change in value of one variable doesn't influence or change the value of other variable. For example, the correlation between weight of person and intelligence is a zero or no correlation.



Zero Correlation

II. Simple, Partial and Multiple Correlation

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

Simple Correlation:

When relation between two variables is studied, it is simple correlation. For example, when one studies relationship between the marks secured by student and the attendance of student in class, it is a problem of simple correlation.

Partial Correlation:

In partial correlation, two or more factors are agreed to be involved but correlation is studied between only two factors, considering other factors to be constant. For example, in above example of relationship between student marks and attendance, the other variable influencing such as effective teaching of teacher, use of teaching aid like computer, smart board etc are assumed to be constant.

Multiple Correlation:

When three or more factors are studied together to find relationships, it is called multiple correlation. For example, in above example if study covers the relationship between student marks, attendance of students, effectiveness of teacher, use of teaching aids etc, it is a case of multiple correlation.

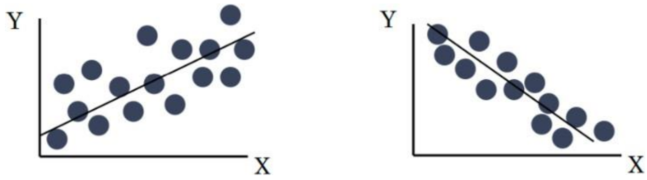
III. Linear and Non-linear Correlation

Depending upon the constancy of the ratio of change between the variables, the correlation may be Linear or Non-linear Correlation.

Linear Correlation:

If the amount of change in one variable bears a constant ratio to the amount of change in the other variable, then correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line.

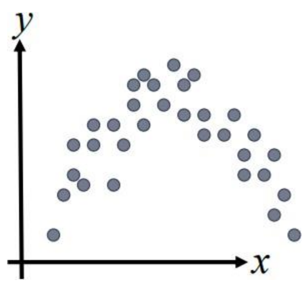
For example: If it is assumed that, to produce one unit of finished product we need 10 units of raw materials, then subsequently to produce 2 units of finished product we need double of the one unit.



Positive linear and Negative linear Correlation

Non-linear Correlation:

If the amount of change in one variable does not bear a constant ratio to the amount of change to the other variable, then correlation is said to be non-linear. If such variables are plotted on a graph, the points would fall on a curve and not on a straight line. For example, if we double the amount of advertisement expenditure, then sales volume would not necessarily be doubled.



Non-linear Correlation

Example:

Given the following pairs of value of the variables X and Y:

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

(a) Make a scatter diagram.

(b) Do you think that there is any correlation between the variables X and Y? Is it positive or negative? Is it high or low?

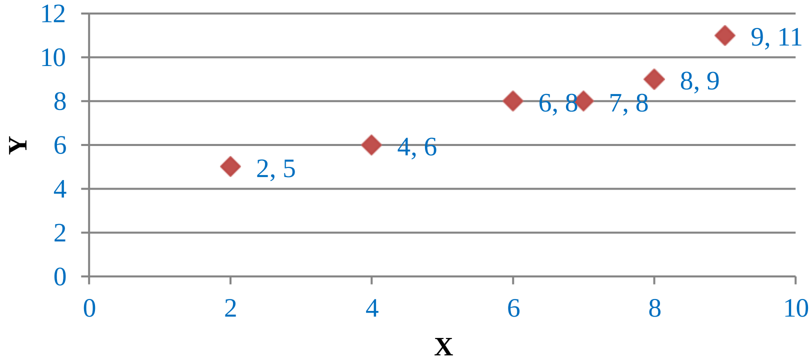


Fig: Scatter Diagram

Karl Pearson's Coefficient of Correlation

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearson's coefficient of correlation, is most widely used in practice. The Pearson's coefficient of correlation is denoted by the symbol r. The formula for computing r is:

$$\mathbf{r} = \frac{\sum \mathbf{xy}}{\mathbf{N\sigma x\sigma y}}$$

Here,

$x = (X - \bar{X})$ and $y = (Y - \bar{Y})$
σ_x = Standard deviation of series X
σ_y = Standard deviation of series Y
N = Number of paired observations.

The above formula for computing Pearson's coefficient of correlation can be transformed in the following form which is easier to apply:

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

Here,

$x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

Characteristics of the correlation coefficient

- A correlation coefficient has no units. The sample correlation coefficient is denoted by r.
- The value of r is always $-1 \le r \le 1$.
- When r = + 1, it means there is ***perfect positive correlation*** between the variables.
- When r = – 1, it means there is ***perfect negative correlation*** between the variables.
- When r > 0 , it means there is ***positive linear correlation*** between the two variables.
- When r < 0, it means there is ***negative linear correlation*** between the two variables.
- When r = 0, it means there is ***no relationship*** between the two variables.

Direct Method of Finding out Correlation Coefficient

Pearson's coefficient of correlation,

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

Here,

$x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

Steps

(i) Take the deviation of X series from the mean $\bar{X}$ and denote the deviations by *x*.
(ii) Square these deviations and obtain the total, *i.e.*, Σx^2.
(iii) Take the deviations of Y series from the mean of Y and denote these deviations by *y*.
(iv) Square these deviations and obtain the total, *i.e.*, Σy^2.
(v) Multiply the deviations x and y, and obtain the total, *i.e.*, Σxy.
(vi) Substitute the values of Σxy , Σx^2 and Σy^2 in the above formula.

Example1

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

X:	6	8	12	15	18	20	24	28	31
Y:	10	12	15	15	18	25	22	26	28

Solution:

X	Y	$x = X-\bar{X}$	$y = Y-\bar{Y}$	x^2	y^2	xy
6	10	-12	-9	144	81	108
8	12	-10	-7	100	49	70
12	15	-6	-4	36	16	24
15	15	-3	-4	9	16	12
18	18	0	-1	0	1	0
20	25	2	6	4	36	12
24	22	6	3	36	9	18
28	26	10	7	100	49	70
31	28	13	9	169	81	117
ΣX=162	ΣY=171			Σx^2= 598	Σy^2=338	Σxy=431

Here;

N= No. of observations = 9

$\bar{X}$= ΣX/N=162/9=18

$\bar{Y}$= ΣY/N=171/9=19

Pearson's coefficient of correlation,

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

$$r= \frac{431}{\sqrt{598 \text{ x } 338}}$$

$$= 0.959$$

Here r=0.959, which is greater than 0. So it is a positive linear correlation.

<u>Example 2</u>

Calculate Karl Pearson's coefficient of correlation from the following data and interpret its value.

Height (inches)	:	62	72	78	58	65	70	66	63	60	72
Weight (Kgs)	:	50	65	63	50	54	60	61	55	54	65

Example 3

Compute the coefficient of correlation between price and demand of ten items

Price (Rs)	:	11	12	13	14	15	16	17	18	19	20
Demand (Rs)	:	30	29	29	25	24	24	24	21	18	15

Solution

Here

N= No. of observations = 10

$\bar{X} = \Sigma X/N = 155/10 = 15.5$

$\bar{Y} = \Sigma Y/N = 239/10 = 23.9$

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	y^2	xy
11	30	-4.5	6.1	20.25	37.21	-27.5
12	29	-3.5	5.1	12.25	26.01	-17.9
13	29	-2.5	5.1	6.25	26.01	-12.8
14	25	-1.5	1.1	2.25	1.21	-1.65
15	24	-0.5	0.1	0.25	0.01	-0.05
16	24	0.5	0.1	0.25	0.01	0.05
17	24	1.5	0.1	2.25	0.01	0.15
18	21	2.5	-2.9	6.25	8.41	-7.25
19	18	3.5	-5.9	12.25	34.81	-20.7
20	15	4.5	-8.9	20.25	79.21	-40.1
155	239			82.5	212.9	-128

Pearson’s coefficient of correlation,

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

$$r = \frac{-128}{\sqrt{82.5 \text{ x } 212.9}}$$

$= -0.9658$

Here r = -0.9658, hence there exists a strong negative correlation between price and demand.

Example 4

Calculate Karl Pearson's coefficient of correlation from the following data and interpret its value.

Roll No of Students	:	1	2	3	4	5
Marks in Accountancy	:	48	35	17	23	47
Marks in Statistics	:	45	20	40	25	45

Example 5

The following table gives indices of industrial production of registered unemployed (in hundred thousand). Calculate the value of the coefficient of correlation so obtained.

Year	1991	1992	1993	1994	1995	1996	1997	1998
Index of production	100	102	104	107	105	112	103	99
No. Unemployed:	15	12	13	11	12	12	19	26

Example 6

Calculate the coefficient of correlation from the following data

X	:	100	200	300	400	500	600	700
Y	:	30	50	60	80	100	110	130

Example 7

Calculate the coefficient of correlation from the following data.

Case	A	B	C	D	E	F	G	H
	10	6	9	10	12	13	11	9
	9	4	6	9	11	13	8	4

Calculation of Correlation Coefficient when Change of Scale and Origin is Made

Since *r* is a pure number, shifting the origin and changing the scale of series do not affect its value.

Example: Find the 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

X	Y	X=X/50	Y= Y/100	x=X-X'	y=Y-Y'	x^2	y^2	xy
300	800	6	8	-4	-4	16	16	16
350	900	7	9	-3	-3	9	9	9
400	1000	8	10	-2	-2	4	4	4
450	1100	9	11	-1	-1	1	1	1
500	1200	10	12	0	0	0	0	0
550	1300	11	13	1	1	1	1	1
600	1400	12	14	2	2	4	4	4
650	1500	13	15	3	3	9	9	9
700	1600	14	16	4	4	16	16	16
		90	108			60	60	60

Here, N=9

X'= 90/9 = 10

Y'= 108/9= 12

Pearson’s coefficient of correlation,

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

r = 60/√(60*60)

= 1.000

REGRESSION

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.

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*

Use of Regression

The study of regression is very useful and important in statistical analysis, which is clear by the following points :

- *Nature of Relationship*: Regression analysis explains the nature of relationship between two variables.
- *Estimation of Relationship*: The mutual relationship between two or more variables can be measured easily by regression analysis.
- *Prediction*: By regression analysis, the value of a dependent variable can be predicated on the basis of the value of an independent variable. For example, if price of a commodity rises, what will be the probable fall in demand, this can be predicted by regression.

Regression Equations

If the variables in a bivariate frequency distribution are correlated, we observe that the points in a scatter diagram cluster around a straight line called the **line of regression.** In a bivariate study, we have **two lines of regression,** namely:

1. Regression of Y on X.
2. Regression of X on Y.

Regression Equation of Y on X

The line of regression of Y on X is used to predict or estimate the value of Y for the given value of the variable X. Thus, Y is the dependent variable and X is an independent variable in this case. The algebraic form of the line of regression of Y on X is of the form:

$$\mathbf{Y = a + bX} \qquad \dots (1)$$

where, a and b are unknown constants to be determined by observed data on the two variables X and Y. Let $(X_1, Y_1), (X_2, Y_2)..., (X_N, Y_N)$ be N pairs of observations on the variable X and Y. Then, for determining a and b in equation (1) we make use of the following normal equations:

$$\Sigma Y = Na + b\Sigma X \qquad \dots (2)$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2 \qquad \dots (3)$$

The values ΣY , ΣX , ΣX^2 and ΣXY can be obtained from the given data.

These normal equations are obtained by minimising the error sum of squares according to the principle of least squares. Solving equations (2) and (3) for a and b, the line of regression of Y on X is completely determined.

Alternatively

There is another way of finding the algebraic form of line of the regression of Y on X. Line of regression of Y on X can also be written in the following form :

$$(Y-\bar{Y}) = r\frac{\sigma Y}{\sigma X}(X-\bar{X})$$

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

Here,

$\bar{Y}$ = the mean of Y

$\bar{X}$ = the mean of X

σY = the S.D. of Y

σX = the S.D. of X

r = the correlation coefficient between X and Y

$b_{YX} = r\frac{\sigma Y}{\sigma X}$ = the regression coefficient of Y on X

From observed bivariate data [(Xi, Yi); i = 1, 2, ... N] the regression coefficient of Y on X, b_{YX} , can be computed from the following formula :

$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

Regression Equation of X on Y

The line of regression of X and Y is used to estimate or predict the value of X for a given value of the variable Y. In this case X is the dependent variable and Y is the independent variable. The standard algebraic form of the line of regression of X on Y is :

X = c + dY ... (4)

where c and d are unknown constant which are determined from the following two normal equations:

$\Sigma X = Nc + d\Sigma Y$ ... (5)

$\Sigma XY = c\Sigma Y + d\Sigma Y^2$ ... (6)

The values of ΣX , ΣY , ΣXY and ΣY^2 can be obtained from observed data. The normal equations (5) and (6) are also obtained by minimising error sum of squares according to the method of least squares. Solving (5) and (6) for c and d and putting these values in (4), the form of regression of X on Y is completely determined.

Alternatively

Like regression of Y on X, the line of regression of X on Y also has an alternative form as

$$\mathbf{(X-\bar{X}) = r\frac{\sigma X}{\sigma Y}(Y-\bar{Y})}$$

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

Where $b_{XY} = r\frac{\sigma X}{\sigma Y}$ is called the *regression coefficient of X on Y*. For observed data, the value of b_{XY} can be computed from the formula:

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

<u>Example 1</u>

From the following data obtain the two regression equations:

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

Solution

X	Y	XY	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
30	**40**	**214**	**220**	**340**

Regression Equation of Y on X

$Y = a + b X$

To determine the value of a and b the following two normal equations are to be solved:

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

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

Substituting the values,

$40 = 5a + 30b$ ... (i)

$214 = 30a + 220b$ ... (ii)

Multiplying Eqn. (i) by 6

$240 = 30a + 180b$ ... (iii)

$214 = 30a + 220b$ ... (iv)

Subtracting Eqn. (iv) from (iii)

$26 = -40b$

$b = -0.65$

Substituting the value of b in Eqn. (i)

$40 = 5a + 30(-0.65)$

$5a = 40 + 19.5 = 59.5$

$a = 11.9$

Putting the values of a and b in the equation, the regression line of Y on X is

$Y = 11.9 - 0.65 X$

Regression Line of X on Y

$X = a + b Y$

and the two normal equations are:

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

$\Sigma XY = a \Sigma Y + b \Sigma Y2$

Substituting the values,

$30 = 5a + 40b$ ... (i)

$214 = 40a + 34b$ ... (ii)

Multiplying Eqn. (i) by 8

$240 = 40a + 320b$ ... (iii)

$214 = 40a + 340b$ ... (iv)

Deducting Eqn. (iv) from (iii)

$-20b = 26$

$\therefore b = -1.3$

Substituting the value of b in Eqn. (i)

$30 = 5a + 40(-1.3)$

$5a = 30 + 52 = 82$

a = 16.4

Putting the values of a and b in the equation, the regression line of X on Y is

$X = 16.4 - 1.3\,Y.$

Example 2:

Obtain regression lines by computing regression coefficients.

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

Solution

X	Y	XY	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
30	**40**	**214**	**220**	**340**

Regression Equation of Y on X

The line of regression of Y on X using its regression coefficient can be written as :

$(Y - \bar{Y}) = b_{YX}(X - \bar{X})$ …(1)

Here,

$\bar{Y} = \Sigma Y/N = 40/5 = 8$

$\bar{X} = \Sigma X/N = 30/5 = 6$

and

$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

$= (5x214-30x40)/(5x220-30^2)$

$= -0.65$

Putting the values of $\bar{X}$, $\bar{Y}$ and b_{YX} in equation (i), one gets the line of regression of Y on X as :

$(Y - 8) = -0.65\,(X - 6)$

or $Y = 8 - 0.65\,X + 3.9$

or $Y + 0.65\,X = 11.9$

$Y = 11.9 - 0.65\,X$

Regression of X on Y

The line of regression of X on Y using its regression coefficient is:

$(X - \bar{X}) = b_{XY}(Y - \bar{Y})$ …(2)

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

$b_{XY} = -1.30$

Putting the value of $\bar{X}$, $\bar{Y}$ and b_{YX} in equation (2), the line of regression of X and Y becomes

$(X - 6) = -1.30\,(Y - 8)$

or $X = 6 - 1.30\,Y + 10.4$

or $X + 1.30\,Y = 16.4$

or $X = 16.4 - 1.30Y$

Example 3

From the following data, obtain the two regression equations

Sales :	91	97	108	121	67	124	51	73	111	57
Purchases:	71	75	69	97	70	91	39	61	80	47

Example 4

From the following data of the age of Husband and the age of Wife, form the two regression equations and calculate the husband's age when the wife's age is 16.

Husband's age :	36	23	27	28	28	29	30	31	33	35
Wife's age	29	18	20	22	27	21	29	27	29	28

Example 5

A panel of two judges P and Q graded seven dramatic performances by independently awarding marks as follows :

Performance	:	1	2	3	4	5	6	7
Marks by P	:	46	42	44	40	43	41	45
Marks by Q	:	40	38	36	35	39	37	41

The eighth performance, for which judge Q could not attend, was awarded 37 marks by judge P. If judge Q has also been present, how many marks would be expected to have awarded by him to the eighth performance.

Properties of the Regression Lines

1. The regression lines of Y on X is used to estimate or predict the best value (in least squares sense) of Y for a given value of the variable X. Here Y is dependent and X is an independent variable.
2. The regression line of X on Y is used to estimate to best value of X for a given value of the variable Y. Here X is dependent and Y is an independent variable.
3. The two lines of regression cut each other at the points $(\bar{X}, \bar{Y})$. Thus, on solving the two lines of regression, we get the values of means of the variables in the bivariate distribution.
4. In a bivariate study, there are two lines of regression. However, in case of perfect correlation that is when r = +1 on –1 we have only one regression line as both the regression lines coincide in this case.
5. When r = 0, i.e., if correlation exists between X and Y, the two lines of regression become perpendicular to each other.