HIGHER SECONDARY COURSE

STATISTICS

CLASS - XII

Government of Kerala
DEPARTMENT OF EDUCATION

State Council of Educational Research and Training (SCERT);
Kerala
2015

Contents

Chapter 1

Correlation Analysis

We know that a bivariate data consists of two variables with a certain relationship. The variables in a bivariate data distribution can be both numerical, both categorical or one numerical and one categorical. Scatter plot is the graphical representation of bivariate data. The degree of variation between two variables is covariance. If there exists some relationship between two variables, and if we study it, then that statistical study is called Bivariate Analysis. Consider the examples-

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the meaning of correlation.
- Recognises different types of correlation.
- Explains the methods of studying correlation.
- Identifies rank correlation coefficient.
- Uses the Karl Pearson's coefficient of correlation.
- Uses rank correlation coefficient in suitable situations.

advertisement cost and sales of a product, price of a commodity and its sale. With increase in the advertisement cost, the quantity sold is bound to increase. Or, with increase in the price of a commodity, the quantity sold is bound to decrease. These relationships may be linear or non linear (curvy linear). Similarly, relationships may exist between two or more variables. In this chapter, we discuss only the linear relationship between two variables.

Correlation analysis is useful in physical and social sciences. It is used to study the relationship between variables. It helps in measuring the degree of relationship and to compare the relationship between variables. Correlation is the basis of the concept of regression which is used for estimation.

1.1 Meaning of Correlation

Correlation refers to the relationship between two variables in a bivariate distribution. We can observe a certain relationship between two variables in the following cases.

- Price of product and its demand.
- Price of product and its supply.
- Wage and price index.
- Height and weight.

In each case, we can statistically analyse the degree or extent to which two variables fluctuate and relate to each other.

Let us consider some cases in detail.

Look at the following cases and corresponding scatter plots.

Case (i): Consider a certain brand of television. The amount utilised for advertisements and quantity sold in different years are given below.

Amount of advertisement (in Rs '000) (X)	25	50	70	80	100	130	170
Quantity sold (Y)	100	220	200	340	370	410	450

Scatter Plot

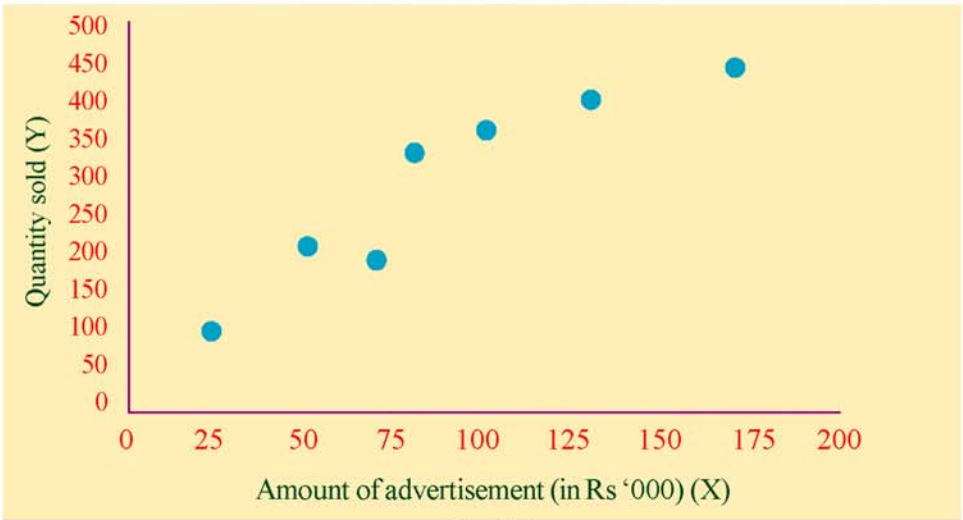


Fig. 1.1

Case (ii): Consider the case of CFL lamps. The price and quantity sold in different months are given below.

Price (X in Rs.)	50	100	120	130	150	200	220	230	240
Quantity sold (Y)	275	234	220	235	160	140	100	130	80

Scatter Plot

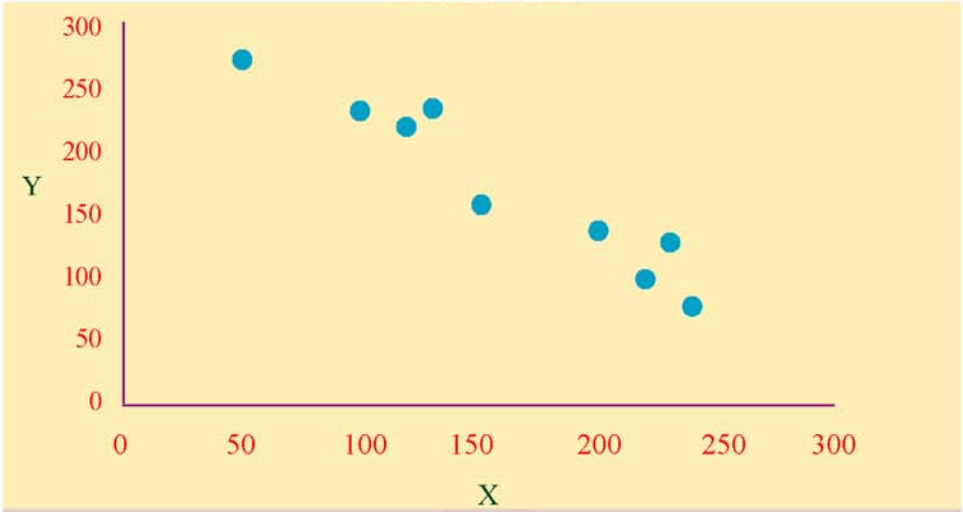


Fig. 1.2

Case (iii): The height in cms. and marks in English out of 50 of 10 students are given as follows.

Height in cms. (X)	150	165	155	156	158	163	158	162	152	167
Marks in English out of 50 (Y)	38	40	35	39	25	27	32	37	28	34

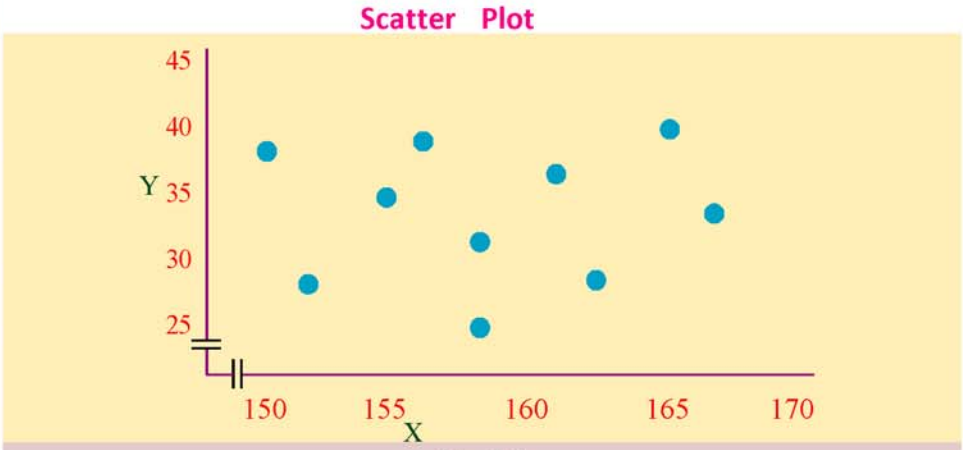


Fig. 1.3

Examine the data and scatter plots in these cases. What is your inference about the relationship between the two variables? We can see that values of X and Y increases together in case (i). while in case (ii), values of Y decreases as the values of X increases and in case (iii), no such relation is seen between X and Y. From this we can conclude that there may be some relation between the variables or there may be no relation between the variables.

Correlation analysis deals with the association or co-variation between two variables and helps to determine the degree of relationship. Correlation is the study of the degree of relationship between two variables. The correlation expresses the relationship or interdependence of two variables upon each other, in such a way that, changes in the values of one variable are sympathetic with the changes in the values of the other. Correlation also shows the degree of co-variation.

> Correlation Analysis is the study of the degree of relationship between two variables in a bivariate distribution.

Know your progress

List any three pairs of related variables which are very familiar to you.

1.2 Types of Correlation

Depending up on the nature of the relationship between the variables, correlation can be classified into:

1. **Positive correlation**
2. **Negative correlation**
3. **No correlation or Zero correlation**

Let us look into some details.

1. Positive Correlation

If the two variables are moving together in the same direction, then the correlation is called positive correlation. That is, increase in the value of one variable is accompanied by an increase in the value of the other variable and decrease in the value of one variable is accompanied by a decrease in the value of the other variable.

Use of fertilizer and yield of crop, price and supply, income and expenditure, etc., are examples for variables with positive correlation.

2. Negative Correlation

If the two variables are moving in opposite direction, then the correlation is called negative correlation. That is, increase in the value of one variable is accompanied by a decrease in the value of the other variable and decrease in the value of one variable is accompanied by an increase in the value of the other variable.

Intensity of light and distance from the source, price and demand, pressure and volume, etc., are examples for variables with negative correlation.

3. No correlation or Zero Correlation

If there is no association between the two variables, we say that there is no correlation or zero correlation. If the change in the value of one variable is not accompanied by any changes in the value of the other variable, then the correlation is zero or the variables have no correlation.

Amount of rainfall and scores in an examination, height and intelligence, etc., are examples for variables with zero correlation.

In some cases the relationship between the two variables may be proportional to each other. This is a case of **perfect correlation.**

Perfect Correlation

If the change in the value of one variable is proportional to the change in the value of the other variable, then the correlation is said to be perfect. If the variables are directely proportional then the correlation is **perfect positive** and if they are inversely proportional, then the correlation is **perfect negative**.

Radius and area of circles, sales and revenue, days of working and income of daily wage workers, hours of working and power consumption of electric appliances are examples for variables with perfect positive correlation.

Examples for variables with perfect negative correlation are pressure and volume (temperature kept constant), speed and time taken for travelling of vehicles, price index and purchasing power of money.

Know your progress

Write examples for the following:

- Positively correlated variables
- Negatively correlated variables
- Perfect Positively correlated variables
- Perfect negatively correlated variables
- Zero correlated variables

1.3 Methods of Studying Correlation

The different methods of studying correlation are discussed below.

1. Scatter Diagram

Scatter diagram is a graphical method of studying correlation. It is the simplest method of finding out whether there is any relationship between the two variables by plotting the values on a chart. It is also known as scatter plot.

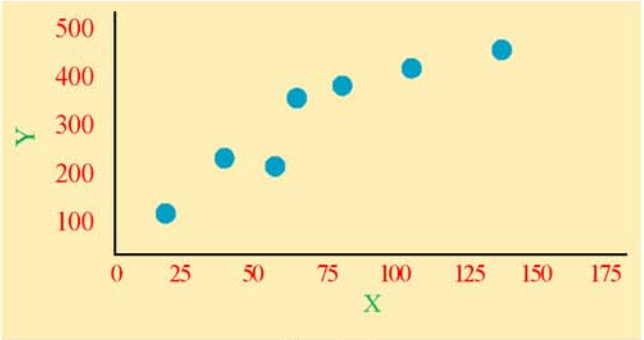


Fig. 1.4

The type of correlation can

be identified by this method.

Look at the scatter plots.

Fig (1.4) reprsents the scatter plot of Price (X) and Supply (Y) of a certain commodity.

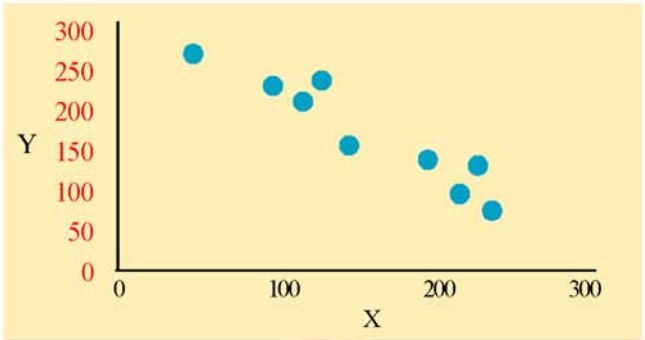


Fig. 1.5

The points in the scatter plot are rising from lower left hand corner to upper right hand corner. It shows that, there is positive correlation between the variables.

Fig (1.5) represents the scatter plot of the Price (X) and Demand (Y) of a commodity.

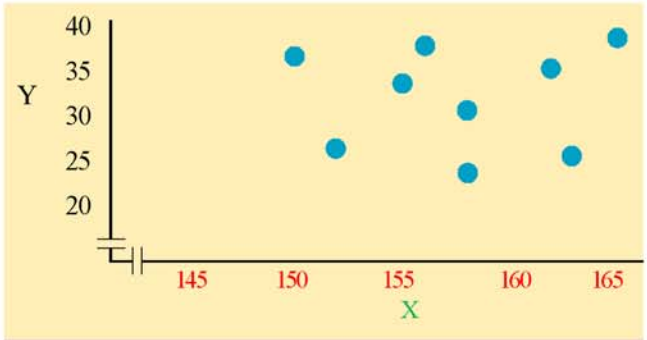


Fig.1.6

The points in the scatter plot are falling from upper left hand corner to lower right hand corner. It shows that, there is negative correlation between the variables.

Fig (1.6) represents the scatter plot of the Height (X) and Scores in Statistics (Y) of students in a group.

The plotted points are scattered all over the diagram. It shows that there is no correlation between the variables.

Fig (1.7) represents the scatter plot of the length of a side (X) and perimeter (Y) of squares.

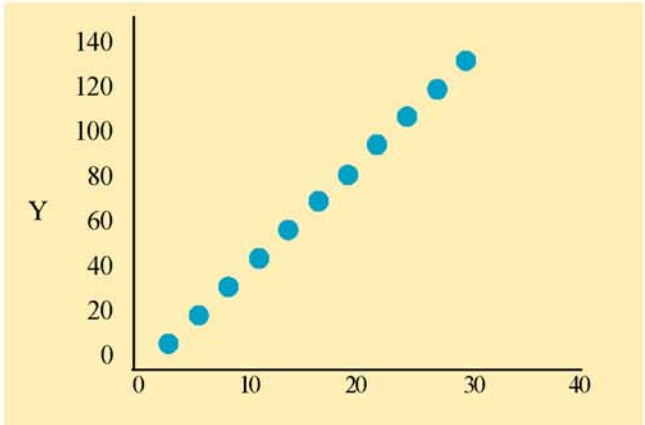


Fig.1.7

The points in the scatter plot are falling in a straight

line from lower left hand corner to upper right hand corner. It shows that, there is perfect positive correlation between the variables.

Fig (1.8) represents the scatter plot of the Age (X) and remaining years of retirement (Y) with regards to teachers in a school.

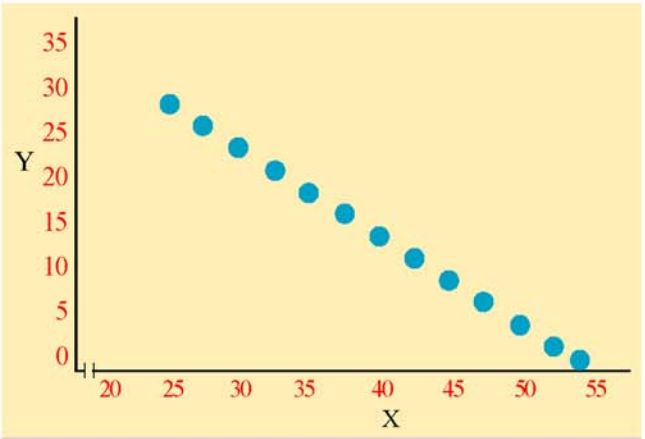


Fig. 1.8

The points in the scatter plot are falling in a stright line from upper left hand corner to lower right hand corner. It shows that, there is perfect negative correlation between the variables.

By observing the scatter diagrams, try to identify the merits and demerits of a scatter diagram. The following are some of them.

Merits

- Simple and attractive
- Easy to understand
- Gives a rough idea at a glance
- Not influenced by extreme items

Demerit

- Does not give the exact degree of correlation

Know your progress

Scores in Economics and Statistics of 10 students in a test are given below. Draw the scatter plot and interpret it.

Scores in Economics out of 100 (X)	85	35	25	14	65	25	78	32	58	45
Scores in Statistics out of 100 (Y)	75	65	32	29	56	18	82	39	62	35

Activity

Collect the scores obtained by 10 students in different subjects in class XI examination and draw the scatter plots for each pair of subjects. Find the subjects which are most correlated, least correlated and not correlated.

Coefficient of Correlation

Coefficient of Correlation is a relative measure showing the degree of relationship between two variables. It is a pure number free from units of measurement which can be used for comparison.

The most commonly used coefficients of correlation are:

- Karl Pearson's Coefficient of Correlation
- Spearman's rank Correlation Coefficient

Karl Pearson's Coefficient of Correlation

Karl Pearson, a great biometrician suggested a mathematical method for measuring the magnitude of the linear relationship between two variables. The most widely used method in practice is Karl Pearson's Coefficient of Correlation. It is usually denoted by 'r'.

Karl Pearson's Coefficient of Correlation between the variables X and Y is given by:

$$r(x,y) = \frac{\text{covariance between x and y}}{(\text{standard deviation of x})(\text{standard deviation of y})}$$

$$= \frac{\text{Cov(x, y)}}{\sigma_x \, \sigma_y}$$

Where

$$\text{cov}(x,y) = \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}$$

$$\therefore\ r = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{1}{n}\Sigma(x-\bar{x})^2}\sqrt{\frac{1}{n}\Sigma(y-\bar{y})^2}}$$

$$r = \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2}\sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}}$$, on simplification

or

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

$$r(x, y) = \frac{\text{Cov(x, y)}}{\sigma_x\ \sigma_y}$$

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

It is obvious that the value of coefficient of correlation (r) always takes values from -1 to +1. That is $-1 \le r \le 1$. This implies that r can be +1, -1, 0, between 0 and +1 and between -1 and 0. Look at the interpretations given below.

Interpretation of Karl Pearson's coefficient of correlation

i. If r = +1, then the correlation is perfect positive.

ii. If r = -1, then the correlation is perfect negative.

iii. If r = 0, then the correlation is zero.

iv. If 0 < r < +1, then the correlation is positive.

v. If -1 < r < 0, then the correlation is negative.

Properties of coefficient of correlation

1) Coefficient of correlation takes any value from -1 to +1.

That is, $-1 \le r \le 1$

2) (i) Magnitude of Coefficient of correlation is unaltered, if a constant is added to (subtract from) the values of one variable (both variables)

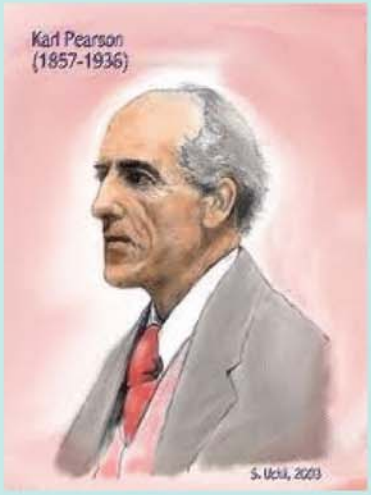
(ii) Magnitude of Coefficient of correlation is unaltered, if the values of one variable (both variables) are multiplied (divided) by a constant.

i.e. $r(u, v) = \pm r\ (x, y)$ where $u = \frac{x - a}{b}$ and $v = \frac{y - c}{d}$; a, b, c and d are constants.

3) Correlation coefficient is symmetric with respect to variables. i.e. $r(x, y) = r(y, x)$.

4) Correlation coefficient between two independent variables is zero. But the converse need not be true.

Karl Pearson was born in London on 27th March 1857. He worked in the University of London and formed the Department of Applied Statistics. He incorporated the biometric and Galton laboratories to this department. He remained with the department until his retirement in 1933 and continued to work till his death in 1936.



Proof for property 2

Let u = $\frac{x-a}{b}$ and v = $\frac{y-c}{d}$

Then Cov (u, v) = $\frac{Cov(x,y)}{bd}$, $\sigma_u = \frac{\sigma_x}{b}$ and $\sigma_v = \frac{\sigma_y}{d}$

$$r(u, v) = \frac{Cov(u,v)}{\sigma_u \sigma_v} = \frac{\frac{Cov(x,y)}{bd}}{\frac{\sigma_x}{b}\frac{\sigma_y}{d}} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = r(x, y)$$

Proof for property 4

The correlation coefficient between the variables given below is zero but the variables are related by the relation $y = (x-20)^2$

X	10	15	20	25	30
Y	100	25	0	25	100

That is, correlation coefficient is zero implies the absence of linear relationship between them. They may however, be related in some other form. Similarly coefficient of correlation may be calculated mathematically from the given values of two variables even though they are really independent.

Illustration 1.1

The following gives the scores obtained in Statistics (X) and Economics (Y) out of 50 by 10 students in class tests.

X	18	25	5	31	12	22	28	23	38	45
Y	16	34	12	28	21	25	31	27	29	40

Draw the scatter diagram and find the Karl Pearson's coefficient of correlation.

Solution:

Scatter diagram

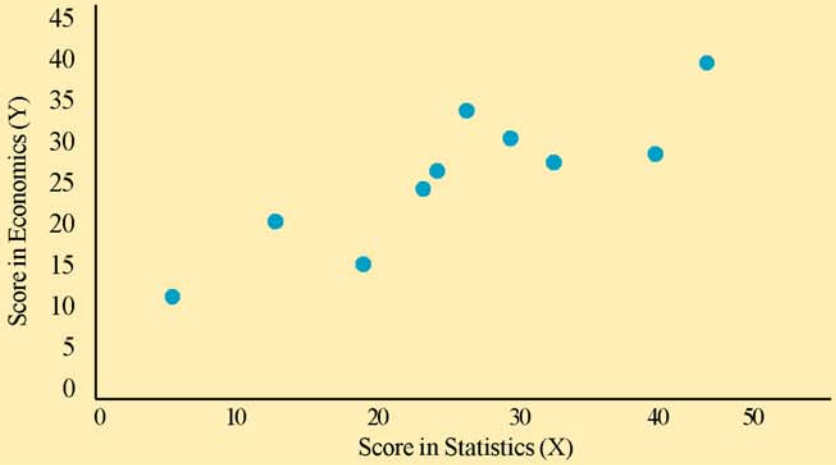


Fig. 1.9

	x	y	x^2	y^2	xy
	18	16	324	256	288
	25	34	625	1156	850
	5	12	25	144	60
	31	28	961	784	868
	12	21	144	441	252
	22	25	484	625	550
	28	31	784	961	868
	23	27	529	729	621
	38	29	1444	841	1102
	45	40	2025	1600	1800
Total	247	263	7345	7537	7259

Karl Pearson's coefficient of correlation, $r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

$$= \frac{10 \times 7259 - 247 \times 263}{\sqrt{10 \times 7345 - (247)^2}\sqrt{10 \times 7537 - (263)^2}}$$

$$= 0.8686$$

The value of the coefficient of correlation 'r' is between 0 and 1. Therefore the correlation between the scores in Statistics and Economics is positive.

Know your progress

Heights (in inches) of 12 fathers (X) and that of their eldest son (Y)are given. Draw the scatter diagram and find the Karl Pearson's coefficient of correlation.

X	65	63	68	69	62	72	70	65	62	64	62	67
Y	66	60	71	67	65	68	63	61	69	62	65	65

Activity

Collect the data on the number of hours of study and scores in a Statistics examination of 20 students in your class. Find the coefficient of correlation and interpret the result.

Spearman's rank correlation coefficient

Some times in a bivariate data, one or both variables are expressed in terms of ranks instead of being expressed in actual values. Generally qualitative characteristics like honesty, beauty, efficiency, intelligence, etc., are better expressed by allotting ranks such as first, second, etc. The study of correlation of the characteristics expressed by ranks is called rank correlation. The primary purpose of computing a correlation coefficient in such a situation is to determine the extent to which the two sets of ranking of some individuals are in agreement or not.

Charles Edward Spearman,

Charles Edward Spearman, a British psychologist found out the method of ascertaining the coefficient of correlation by ranks. This measure is useful in dealing with qualitative characteristics.

Spearman's rank correlation coefficient is denoted by ρ and is given by:

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} \text{ or } \rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

Where d = difference of ranks of an individual and n = number of individuals.

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)} \text{ or } \rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

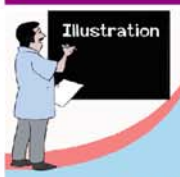


Illustration 1.2

Ranks obtained by 10 students in a Mathematics examination and their ranks in an intelligence test is given below.

Ranks in Mathematics	1	3	6	10	5	9	2	4	7	8
Ranks in intelligence test	4	5	3	8	9	10	1	2	7	6

Find the rank correlation coefficient.

Solution:

Here the ranks of the students are given.

Ranks in Mathematics	1	3	6	10	5	9	2	4	7	8
Ranks in intelligence test	4	5	3	8	9	10	1	2	7	6

X	Y	d	d^2
1	4	-3	9
3	5	-2	4
6	3	3	9
10	8	2	4
5	9	-4	16
9	10	-1	1
2	1	1	1
4	2	2	4
7	7	0	0
8	6	2	4
			Total: 52

$$\rho = 1 - \frac{6\Sigma d^2}{n^3 - n}$$

$$= 1 - \frac{6 \times 52}{10^3 - 10}$$

$$= 1 - \frac{312}{990}$$

$$= 1 - 0.3152$$

$$= 0.6848$$

Know your progress

The ranks given by two judges to 10 competitors in a beauty contest are as follows. Find the rank correlation coefficient.

Judge 1	3	5	8	6	4	9	7	2	1	10
Judge 2	2	6	10	8	3	7	5	1	4	9

Calculation of rank correlation coefficient when the ranks are repeated

Sometimes it may be necessary to assign equal ranks to two or more items. In such cases, it is customary to give an average rank to each item. Thus, if two items are ranked equal say at second place, each of them can be given the rank $\frac{2+3}{2}$, that is 2.5. Similarly if three items are ranked equal say at fifth place, each of them can be given the rank $\frac{5+6+7}{3}$, that is 6. When equal ranks are assigned to some entries a correction factor is to be added to the value of Σd^2 in the above formula for calculating the rank coefficient of correlation. The Correction Factor (C.F) is given by $\frac{\Sigma(m^3 - m)}{12}$, where m stands for the number of items with common rank. If there are more than one such group of items with common rank, this value is added as many times as the number of such group. The formula can be written as: $\rho = 1 - \frac{6\Sigma d^2 + C.F}{n^3 - n}$

Where C. F $= \frac{\Sigma(m^3 - m)}{12}$

Illustration 1.3

A competitive test includes written test, group discussion and interview. The scores obtained in the written test by 10 top rank holders are given below.

Rank	1	2	3	4	5	6	7	8	9	10
Scores in written test	78	63	65	62	63	58	63	52	50	52

Find the rank correlation coefficient between the final rank and scores in the written test.

Solution:

First we have to rank the individuals according to the scores in the written test. The score 63 is repeated 3 times in the third, fourth and fifth places. Similarly 52 is repeated twice in the places eighth and ninth. Therefore rank 4 is assigned to the persons scored 63 and rank 8.5 is assigned to persons with score 52.

Rank at final	1	2	3	4	5	6	7	8	9	10
Rank in written test	1	4	2	5	4	7	4	8.5	10	8.5

R_1	R_2	d	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	4	1	1
6	7	-1	1
7	4	3	9
8	8.5	-0.5	0.25
9	10	-1	1
10	8.5	1.5	2.25
		Total	**20.5**

Rank 4 is repeated three times.

There for C. F $= \frac{1}{12}(m^3 - m) = \frac{1}{12}(3^3 - 3) = 2$

Rank 8.5 is repeated two times.

There for C. F $= \frac{1}{12}(m^3 - m) = \frac{1}{12}(2^3 - 2) = 0.5$

Total C.F = 2+0.5 = 2.5

$$\rho = 1 - \frac{6\Sigma d^2 + C.F}{n^3 - n}$$

$$= 1 - \frac{6 \times 20.5 + 2.5}{10^3 - 10}$$

$$= 1 - 0.1393$$

$$= \mathbf{0.8607}$$

Know your progress

The scores obtained by 12 students in a written test and ranks in performance are given below. Find rank correlation coefficient.

Scores in written test	12	15	18	20	18	16	13	18	17	11	13	18
Rank in performance	8	7	2	1	6	5	11	4	9	12	10	3

Activity

Conduct a quiz competition based on Statistics. Find the rank correlation between the ranks of the top 10 students in the quiz competition and their scores in statistics class test.

Let us conclude

Correlation is the study of relationship between two variables. Correlation can be studied graphically using scatter diagram. Different types of correlation are positive, negative and no correlation. If the variables are directly proportional, then the correlation is perfect positive and if the variables are inversely proportional, then the correlation is perfect negative. Correlation coefficient is the measure of degree of relationship between two variables. Karl Pearson's coefficient of correlation is most widely used. If one of the variables or both are qualitative, then we use Spearman's rank correlation coefficient to find the degree of relationship. The method of finding rank correlation coefficient when the ranks are repeated is also explained in this chapter.

Lab Activity

Verify the results obtained in Illustration 1.3 using spread sheet application.

Let us assess

For Questions 1-6, choose the correct answer from the given choices.

1. The maximum value of coefficient of correlation is:

   a) 0 b) 1 c) -1 d) infinity

2. The value of the coefficient of correlation:

   a) Has no limit b) Can be greater than 1

   c) Can be less than -1 d) Varies from -1 to +1

3. The coefficient of correlation will be:

   a) Positive b) Negative

   c) Either positive or negative d) Positive or negtive or zero

4. If the correlation between the variables X and Y is 0.3, then the correlation between the variables 2X and Y is.....

   a) 0.6 b) 0.9 c) 0.3 d) 0.4

5. If the value of Pearson's correlation coefficient calculated for marks in Statistics and Economics of 100 students is 0.9, there exists ..... type of correlation between the two variables.

   a) High positive b) High negative

   c) Perfect positive d) Perfect negative

6. If the correlation coefficient between the two variables X and Y is 0.4, then the correlation coefficient between X+3 and Y-5 is .....

   a) 0.8 b) 0.4 c) 0.2 d) 0.6

7. Raw cotton imports and cotton manufacture in million tons of a certain state for different years are given. Construct a scatter diagram.

Raw cotton imports (in million tons)	47	64	100	97	126	203	170	115
Cotton manufacture (in million tons)	70	85	100	103	111	139	133	115

8. The price in rupees(X) and supply in quintals (Y) of biriyani rice in a whole sale store is given. Draw a scatter diagram and interpret it.

X	10	22	34	35	69	85	95	98
Y	32	36	25	45	32	56	86	68

9. Calculate Karl Pearson's coefficient of correlation between price and supply of commodity of a retail dealer from the following data.

Price (in Rs.)	25	38	29	32	35	38	40	42
Supply (in Kg.)	38	35	39	45	42	48	39	52

10. Calculate the coefficient of correlation between the height of fathers and sons from the data given below.

Height of father (in inches)	64	65	66	67	68	69	70
Height of son (in inches)	66	67	65	68	70	68	72

11. The following data relate to the income (x) and expenditure (y) of 5 workers. Compute Pearson's correlation coefficient.

$\Sigma(x-\bar{x})^2 = 1000,\ \Sigma(y-\bar{y})^2 = 40,\ \Sigma(x-\bar{x})(y-\bar{y}) = 100$

12. The covariance between the variables X and Y is 10 and the variances of X and Y are 16 and 9 respectively. Find the coefficient of correlation.

13. Calculate the coefficient of correlation between age of cars and annual maintenance cost and comment.

Age of cars (years) X	2	4	6	7	8	10	12
Annual maintenance cost (rupees) Y	16000	15000	18000	19000	17000	21000	20000

14. The ranks of 10 students in two subjects of an examination is given as follows

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

Find rank correlation coefficient.

15. The marks in XI and XII examinations for 5 students in Statistics are given below. Compute the rank correlation coefficient.

Marks in XI (x) :	32	49	50	28	30
Marks in XII (y) :	40	50	55	25	43

16. The scores given by two judges in an elocution competition for 5 competitors are as follows:

Judge I	70	65	72	64	78
Judge II	91	76	66	48	55

Find the rank correlation coefficient.

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

x	:	22	24	27	35	21	20	27	25	27	23
y	:	30	38	40	50	38	25	38	36	41	32

18. Find the spearman's rank coefficient of correlation between sales and profits of the following 10 firms

Firms :	A	B	C	D	E	F	G	H	I	J
Sales :	50	50	55	60	65	65	65	60	60	50
Profit :	11	13	14	16	16	15	15	14	13	13

19. The marks in Class XI and the Class XII exams for 7 higher secondary students in Statistics are given below. Compute the rank correlation.

Marks in Class XI :	15	14	25	14	14	20	22
Marks in Class XII :	25	12	18	25	40	10	7

Chapter 2

Regression Analysis

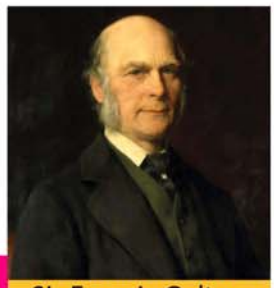
The correlation coefficient we have discussed in the previous chapter simply tells us about the direction and strength of relationship between two variables. In 1889 Sir Francis Galton published a paper on heredity. He reported his findings based on the study of relationship between the heights of fathers and their sons. He observed that the height of offsprings regress towards the mean. While dealing with economic and commerce data, we are required to make prediction and estimation. Prediction is one of the major problems in almost all spheres of human activity. Regression

Significant Learning Outcomes

After the completion of this chapter, the learner:

- Identifies the concept of regression analysis.
- Estimates unknown values for corresponding values given.
- Recognises regression lines and their point of intersection.
- Explains properties of regression coefficients.
- Compares correlation and regression.

analysis is one of the scientific techniques for making such prediction. Regression analysis also measures the percentage variation in dependent variable due to the influence of independent variable. It is one of the most widely used statistical techniques in almost all real life situations. We study more about regression analysis in this chapter.


Sir Francis Galton

2.1 Meaning of regression

Regression is a measure of the functional form of relationship between the variables. Or, in other words, it is a mathematical measure of the nature of relationship between the variables. The word regression means 'going back'. It is the study of cause and effect relationship. In this chapter we will focus only on linear regression, which involves only two variables. They are dependent variable and independent variable. It helps us to estimate the unknown values of dependent variables from the known values of independent variables.

For e.g. : By using the technique of regression, an economist may be able to estimate the demand of a commodity for a given price or an agriculturist can predict production based on rainfall.

> Regression analysis is a mathematical measure of the nature of relationship between two or more variables.

Independent variable and dependent variable

Suppose, a researcher studies the effect of age on a person's blood pressure. Here 'age' is an independent variable and 'blood pressure' is a dependent variable. If expenditure of a person depends on his income, the variable 'income' is independent variable and 'expenditure' is dependent variable. The variable whose value is to be predicted, is called dependent or response variable and the variable used for prediction is called independent or predictor variable.

2.2 Linear regression

When the given bivariate data are plotted on a graph paper we get a scatter diagram. We can construct straight lines through the points in the scatter diagram as shown in the figure given below.

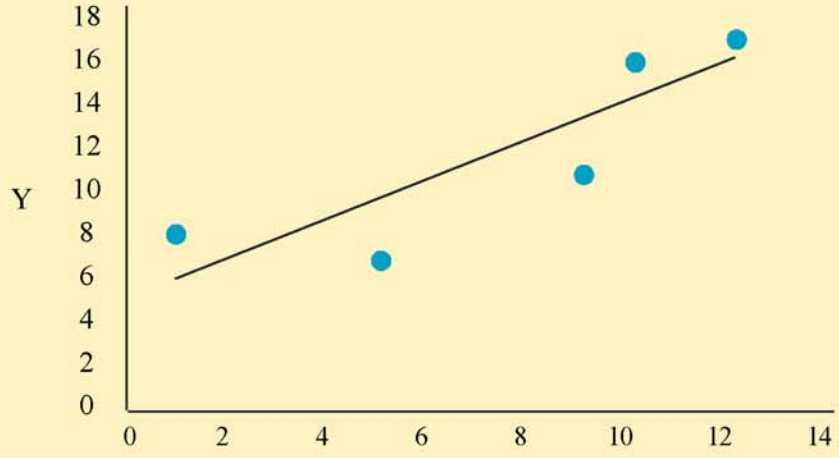


Fig. 2.1

If the points on the scatter diagram concentrate around a straight line that line is called **regression line** or **line of best fit.** The line of best fit is that line which is closer to the points in the scatter diagram. The equation of such a line is the first degree equation in X and Y. Since the relationship between the variables X and Y is not reversible we have two regression lines. One regression line shows regression equation of Y on X and other shows regression equation of X on Y.

Regression line of Y on X is used to predict Y for a given value of X and Regression line of X on Y is used to predict X for a given value of Y.

2.3 Regression equations

Regression equations are the equations of the regression lines. When we have two variables X and Y, we can think of two regression lines. One is regression equation of Y on X and other is regression equation of X on Y. Regression equations can be derived using Legendre's principle of least squares. In regression equation of Y on X, Y is dependent variable and X is independent variable.

Regression equation of Y on X is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Where $\bar{y} = \frac{\Sigma y}{n}$ = mean value of Y

$\bar{x} = \frac{\Sigma x}{n}$ = mean value of X

b_{yx} = Regression coefficient of Y on X

$= \frac{Cov(X, Y)}{\text{var } X}$ or $b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$

Similarly when X is dependent variable and Y is independent variable we have another equation known as regression equation of X on Y. It is obtained by the formula.

Regression equation of X on Y

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

Where b_{xy} is the regression coefficient of X on Y.

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

Principle of least squares states that sum of squares of vertical deviations from the observed values and values obtained by the line of best fit should be minimum. i.e., if $d_1, d_2, d_3 \ldots\ldots$ are the deviations then principle of least squares states that the line of best fit should be drawn so as $d_1^2 + d_2^2 + d_3^2 + \ldots\ldots$ is minimum.

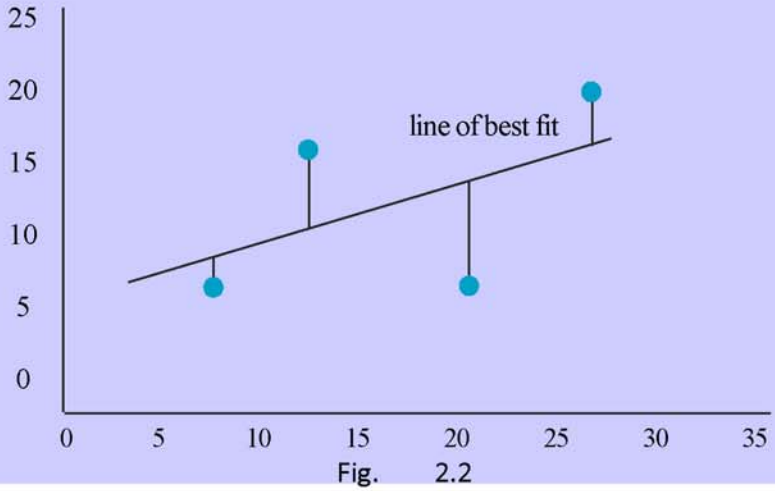


Fig. 2.2

Illustration 2.1

The following data relate to sales and purchase of 10 important shops in a city.

Sales (000's):	4	6	5	9	10	7	2
Purchase(000's):	2	5	3	7	7	3	1

Form the regression equation and also evaluate the amount of sales for a purchase of Rs. 9000.

Solution

Let us take sales as the variable X and purchase as the variable Y

X	Y	XY	Y^2
4	2	8	4
6	5	30	25
5	3	15	9
9	7	63	49
10	7	70	49
7	3	21	9
2	1	2	1
$\Sigma x = 43$	$\Sigma y = 28$	$\Sigma XY = 209$	$\Sigma Y^2 = 146$

The equation of regression line of X on Y is

$$\bar{y} = \frac{\Sigma y}{n} = \frac{28}{7} = 4 \qquad \bar{x} = \frac{\Sigma x}{n} = \frac{43}{7} = 6.14$$

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

$$= \frac{7 \times 209 - 43 \times 28}{7 \times 46 - (28)^2}$$

$$= 1.08$$

The equation is $x - \bar{x} = b_{xy}(y - \bar{y})$

i.e., $x - 6.14 = 1.08\,(y - 4)$

To find the amount of sales when purchase is 9000, we put Y = 9 in the above regression equation

$$x - 6.14 = 1.08\,(9 - 4)$$
$$x = 6.14 + 1.08 \times 5$$
$$= 11.54$$

Know your progress

The following data relate to the experience of machine operators and their performance ratings.

Operator experience(X) in years	16	12	18	4	3	10	5	12
Performance ratings(Y)	87	88	89	68	78	80	75	83

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

Properties of regression coefficients

The following are important properties of regression coefficients.

- In regression equation of y on x , b_{yx} is the coefficient of X.
- In regression equation of x on y, b_{xy} is the coefficient of Y.

  E.g.: $y - 4 = 1.2\,(x - 2)$, $b_{yx} = 1.2$

  $x - 6 = 0.7\,(y - 2)$, $b_{xy} = 0.7$
- The signs of both regression coefficients are same. That is, regression coefficients are either both positive or both negative.
- The product of both regression coefficients should be below one. That is, $b_{yx} \cdot b_{xy} \leq 1$.
- The geometric mean of the regression coefficient is Coefficient of correlation i.e., $r = \pm\sqrt{b_{yx} \times b_{xy}}$
- $b_{yx} = r \times \frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r \times \frac{\sigma_x}{\sigma_y}$ where σ_y is standard deviation of Y and σ_x is standard deviation of X.
- $b_{yx} \neq b_{xy}$ in general.

$$b_{yx} \times b_{xy} = r \times \frac{\sigma_y}{\sigma_x} \times r \times \frac{\sigma_x}{\sigma_y} = r^2,$$

i.e., $r = \pm\sqrt{b_{yx} \times b_{xy}}$, where r is correlation coefficient.

When both b_{yx} and b_{xy} are positive, r is positive.

When both b_{yx} and b_{xy} are negative, r is negative.

Illustration 2.2

The following data relate to area of cultivation in hectores of land(X) and agricultural output in tonnes(Y).

	X	Y
Arithmetic mean	50	30
Standard deviation	5	2

Coefficient of correlation = 0.7

1) Calculate the regression equation of agricultural output on area of cultivation.
2) Estimate agricultural output when there are 80 hectores of land available.

Solution

Given $\bar{y} = 30$ $\sigma_y = 2$

$\bar{x} = 50$ $\sigma_x = 5$

r = 0.7

1) To find the regression equation of agricultural output on area of cultivation, we need to find the Regression equation of Y on X

Regression coefficient of Y on X, $(b_{yx}) = r \times \frac{\sigma y}{\sigma x}$

$$b_{yx} = 0.7 \times \frac{2}{5} = 0.28$$

Regression equation of Y on X is $y - \bar{y} = b_{yx}(x - \bar{x})$

$$y - 30 = 0.28\,(x - 50)$$

2) To estimate agricultural ouput on Area of cultivation, substitute x = 80

$$y - 30 = 0.28\ (80 - 50)$$

$$y - 30 = 0.28\ (30) = 8.4$$

$$y = 30 + 8.4 = 38.4$$

Illustration 2.3

Let $4x + 5y - 10 = 0$ is the regression equation of Y on X. Find b_{yx}.

Solution

$$5y = 4x + 10$$

$$\text{i.e., } y = \frac{4}{5}x + \frac{10}{5}$$

We know that in a regression equation of y on x, b_{yx} is the coefficient of x.

$$\therefore\ b_{yx} = \frac{4}{5}, \text{ the coefficient of x.}$$

Illustration 2.4

Calculate correlation coefficient if $b_{yx} = -0.23$ and $b_{xy} = -0.75$.

Solution

$$r = \pm\sqrt{b_{yx}b_{xy}} = \pm\sqrt{(-0.23)(-0.75)}$$

$$= \pm\sqrt{.1725} = \pm 0.4153$$

$r = -0.4153$ (Since b_{yx} and b_{xy} are negative.)

Know your progress

1. The following data are given for marks in English and Statistics in a certain examination.

	English	Statistics
Mean Marks	39.5	47.5
S.D of Marks	10.8	16.8

Correlation coefficient between Marks in Statistics & English = 0.42.

(a) Find the most probable mark in English if marks in Statistics is 50.

(b) Estimate the marks in Statistics if Marks in English is 35.

2. Regression coefficient between X on Y is $\frac{9}{16}$ Correlation coefficient between the same variables is $\frac{1}{4}$. Find Regression coefficient between X on Y.

Identification of regression lines

Regression lines are not reversible. Therefore when we have two regression lines an important problem is to identify which one is regression equation of Y on X and which one is regression equation of X on Y. By supposing one of the equation as the regression equation of Y on X and other as X on Y, we can obtain regression coefficients. If the product of the these two is numerically less than one then our supposition is true. However if their product is greater than one then our supposition is wrong. The regression lines can be identified as in the following example.

Illustration 2.5

Out of the two lines of regression given by $x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ which one is the regression line of X on Y and which one is regression line of Y on X.

Solution:

$x + 2y - 5 = 0$.....(1)

$2x + 3y - 8 = 0$.....(2)

Let us assume equation (1) as the regression equation of X on Y and equation (2) as regression equation of Y on X.

Let equation (1) be written as $x = -2y + 5$ therefore $b_{xy} = -2$

Equation (2) as $y = \frac{-2}{3}x + \frac{8}{3}$ therefore $byx = \frac{-2}{3}$

$$b_{xy} \times b_{yx} = -2 \times \frac{-2}{3} = \frac{4}{3} = 1.33 \text{ which is greater than } 1$$

$\therefore$ our supposition is wrong . That means (1) is regression equation of Y on X and (2) is regression equation of X on Y.

Know your progress

For a group of 50 persons, the regression equations of age (X) and the blood pressure (Y) are $3y - 5x + 180 = 0$ and $4x + 10y + 100 = 0$. Find the correlation coefficient.

Point of intersection of two regression lines

When we have two regression lines, they coincide at the point $(\bar{x}, \bar{y})$ as given in the following figure

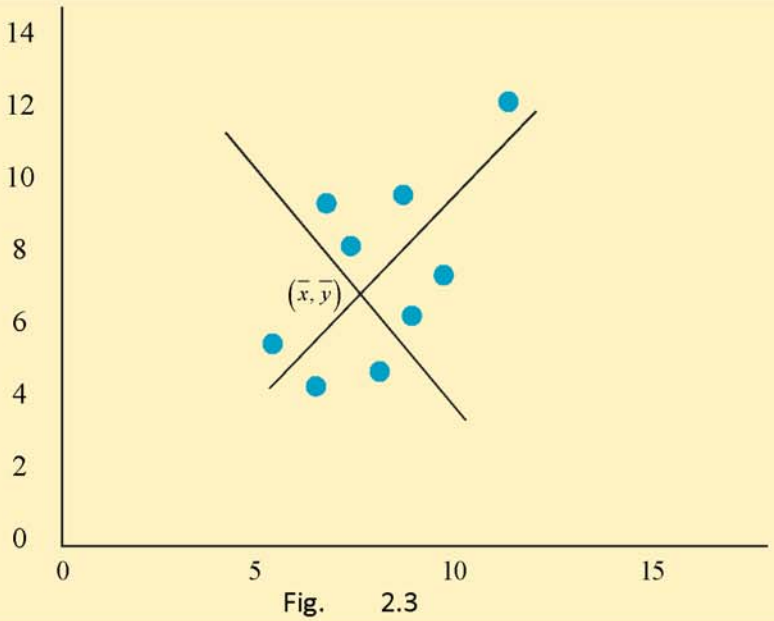


Fig. 2.3

When r = 1, the two regression lines coincide

When r = 0, the two regression lines are perpendicular

Illustration 2.4

In a linear regression analysis of 20 observations, the two lines of regression are $10y + 7x - 4 = 0$ and $5x + 9y - 1 = 0$.

a) Identify regression lines.

b) Obtain the correlation coefficient.

c) Obtain the mean values of X and Y.

Solution

a) $10y + 7x - 4 = 0$.....(1)

$5x+9y - 1 = 0$.....(2)

Let us assume equation (1) as the regression equation of X on Y and equation (2) as regression equation of of Y on X.

Equation (1) can be written as $7x = -10y+4$. Therefore $b_{xy} = -\frac{10}{7}$

Equation (2) can be written as $9y = -5x+1$. Therefore $b_{yx} = -\frac{5}{9}$

$$b_{xy} \times b_{yx} = -\frac{10}{7} \times -\frac{5}{9}$$

$= 0.79$ which is less than one.

$\therefore$ Our supposition is correct . That means (1) is regression equation of Y on X and (2) is regression equation of X on Y.

b) $r = \pm\sqrt{b_{xy} \times b_{yx}} = \pm\sqrt{-\frac{10}{7} \times -\frac{5}{9}}$

$= \pm\sqrt{0.79} = \pm 0.89$

$r = -0.89$ (Since b_{yx} and b_{xy} are negative)

c) Since both the lines of regression pass through the mean values, the point $(\bar{X}, \bar{Y})$ will satisfy both the equations. Hence both the equations can be written as

$7\bar{x} + 10\bar{y} = 4$ .....(1)

$5\bar{x} + 9\bar{y} = 1$ .....(2)

$(1) \times 5 \rightarrow \quad 35\bar{x} + 50\bar{y} = 20$.....(3)

$(2) \times 7 \rightarrow \quad 35\bar{x} + 63\bar{y} = 7$.....(4)

Substracting equation (3) from (4) we get $13\bar{y} = -13$

$$\therefore \bar{y} = -1$$

Putting the value of $\bar{y} = -1$ in equation (1) we get $\bar{x} = +2$

Thus mean of X = 2 and mean of Y = –1

Comparison between correlation and regression

Regression	Correlation
1. Regression is asymmetric.	Correlation is symmetric.
2. Regression is the cause and effect relationship between the variables.	Correlation is the association between the variables.
3. It is used for prediction.	It is not used for prediction.
4. Regression is the study of the nature of relationship between the variables.	Correlation is the study of strength of relationship.
5. It is used for further mathematical treatment.	It is not used for further mathematical treatment.
6. Regression is not reversible.	Correlation is reversible.

Let us conclude

In this chapter we have discussed the concept and importance of measuring regression. Regression is widely used for prediction and forecasting . The knowledge of regression helps us to understand how the value of dependent variable changes when the value of independent variable is fixed. We have also discussed about two regression lines and their importance. The comparison between correlation and regression will give us the characteristics of correlation and regression.

Lab Activity

(1) The following are the weights (Kg.) and blood glucose levels (mg./100ml.)of 16 apparently healthy adult males.

Weight	64	75	73	82	76	95	76	82
Glucose	108	109	104	102	105	121	99	100

(a) Obtain the linear regression equations

(b) Predict the glucose level of a person who weighs 95 Kg.

2 The body weight and the Body Mass Index (BMI) of 7 school going children are given in the following table.

Weight (Kg.):	15.0	26.0	27.0	25.0	25.5	27.0	32.0
BMI:	13.35	16.12	16.74	16.00	13.59	15.73	15.65

(a) Find the regression equation of BMI with respect to weight.

(b) Estimate BMI when weight is 40kg.

Let us assess

For Questions 1-10, choose the correct answer from the given choices.

1. If $b_{yx} \geq 1$ then b_{xy} is .....
   a) less than 1 b) greater than 1 c) equal to 1 d) equal to -1
2. The term regression was introduced by ……..
   a) R A Fisher b) Sir Francis Galton
   c) Karl Pearson d) none of these
3. If X and Y are two variables, then there can be at the most ..... number of regression lines.
   a) one b) two c) three d) infinite
4. If the correlation coefficient between two variables X and Y is negative , then regression coefficient Y on X is ……..
   a) positive b) negative c) zero d) not certain
5. In a regression line of Y on X , the variable X is known as …….
   a) independent variable b) regressor
   c) explanatory variable c) all of the above
6. The geometric mean of two regression coefficients b_{yx} and b_{xy} is equal to .....
   a) r b) r^2 c) 1 d) none of the above
7. Which one of the following can be regression coefficients:
   a) $\left(1, \frac{3}{2}\right)$ b) $\left(\frac{1}{2}, \frac{3}{2}\right)$ c) $\left(2, \frac{3}{2}\right)$ d) (2, 3)
8. Let $2x + 3y - 5 = 0$ is the regression line of X on Y then b_{xy}=.....
   a) $\frac{3}{2}$ b) $\frac{2}{3}$ c) $\frac{-2}{3}$ d) $\frac{-3}{2}$

9. Let $b_{yx} = -0.5$, $b_{xy} = -0.3$ then the value correlation coefficient=.....

   a) –0.15 b) 0.15 c) 0.39 d) –0.39

10. To estimate the value of Y for a given value of X ,the regression equation used is .....

    a) Y on X b) X on Y c) both of these d) none of these.

11. The following data relate to the age of drivers(X) the and number of motor accidents which occurred in a locality (Y) during the last 6 months.

    a) Form the suitable regression equation.

    b) Using the above equation calculate the number of accidents caused by a of 20 year old person.

Age of drivers(years)	19	21	30	45	50	54	25
Number of motor accidents	50	52	40	22	10	14	35

12. Given is the data on price and sales of a particular commodity.

Price	20	25	50	15	25	30	20	17
Sales	15	11	10	30	15	17	20	12

Using the technique of regression evaluate sales when the price of the commodity is 40 rupees.

13. The following data relate to time spent for exercising daily in minutes(X) and blood pressure(Y) of a group of patients.

	X	Y
Mean	60	100
Standard Deviation	20	15

Correlation coefficient = –0.81

a. Find the equation of suitable regression line.

b. Calculate the blood pressure of a person who exercised 70 minutes daily.

14. The following data relate to advertising expenditure and sales of 10 major shops in a city.

	Advertising expenditure (lakhs)	Sales (lakhs)
Mean	10	15
SD	5	3

Coefficient of correlation = 0.65

a) Calculate sales when advertising expenditure is 13 lakhs.

b) Calculate advertising expenditure when sales is 20 lakhs.

15. The following calculations have been made for the price of 12 stocks (X) on BSE on a certain day along with volume of sales on shares (Y). From these calculations calculate the regression equation of price of stocks on volume of shares.

$\sum x = 580,\ \sum y = 370,\ \sum xy = 11494,\ \sum x^2 = 41658,\ \sum y^2 = 17206$

16. While studying about the relationship between scores on statistics (X) and scores on accountancy (Y) the following regression equations were obtained.

Regression equation of Y on x: $3y - 2x - 100 = 0$

Regression equation of X on Y: $4y - 3x + 50 = 0$

a) Find the correlation coefficient.

b) Estimate the scores on Statistics if a student got 50 score in Accountancy.

17. Find the mean values of the variables X and Y for the following regression equations.

Regression line of Y on X : $2y - x = 50$

Regression line of X on Y : $3y - 2x = 10$

18. The following regression equations are obtained when studying about the demand (X) and supply(Y) of a group of commodities.

$26 - 3x - 2y = 0$

$31 - 6x - y = 0$

a) Identify regression lines and find the correlation coefficient.

b) Find the mean values of the variables X and Y.

19. In the study of regression lines, regression coefficient of Y on X = 0.75, correlation coefficient =0.5, standard deviation of Y= 4. Find SD of X.

20. In the study of regression lines, regression coefficient of X on Y = $\frac{3}{5}$, variance of Y=30, correlation coefficient = $\frac{5}{6}$. Find variance of X.

21. In a regression analysis of the income tax of government employees in thousands (X) and their annual income in lakhs (Y), the following regression equations have been obtained.

    $25x - 10y + 10 = 0$

    $10y - 7x - 100 = 0$

    a) Identify regression lines and hence find the correlation coefficient.

    b) Find the mean values of the variables X and Y.

    c) If variance of X = 36, find the variance of Y.

22. A regression analysis on the income in thousands (Y) and expenditure in thousands (X) resulted in the following regression equations.

    $x - y - 3 = 0$ and $5x - 8y + 15 = 0$

    a) Identify regression lines.

    b) What is the correlation coefficient between income and expenditure?

    c) Find the values of $\overline{X}$ and $\overline{Y}$.

    d) What is the most probable value of income when expenditure is 2000?

23. If two lines of regression are $4x - 5y + 10 = 0$ and $20x - 9y - 75 = 0$:

    a) Which of these lines is regression equation of X on Y?

    b) Find correlation coefficient.

    c) Find standard deviation of y if standard deviation of X = 5.

24. The equation of two regression lines between two variables are expressed as $2x - 3y = 0$ and $4y - 5x - 7 = 0$.

    a) Identify which of two can be called regression line of Y on X and X on Y.

    b) Find the correlation coefficient.

    c) Find mean value of X and mean value of Y.