Chapter 9

CORRELATION AND REGRESSION ANALYSIS

Learning Objectives

After studying this chapter students will be able to understand

- Concept of Karl Pearson's correlation co-efficient and the methods of computing it.
- Spearman's Rank correlation co-efficient
- Concept of Regression and Regression co-efficients
- Regression lines both x on y and y on x.

9.1 Correlation

Introduction

In the previous Chapter we have studied the characteristics of only one variable; example, marks, weights, heights, rainfalls, prices, ages, sales, etc. This type of analysis is called univariate analysis. Sometimes we may be interested to find if there is any relationship between the two variables under study. For example, the price of the commodity and its sale, height of a father and height of his son, price and demand, yield and rainfall, height and weight and so on. Thus the association of any two variables is known as correlation.

Karl Pearson

IKYFAG

Correlation is the statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.

9.1.1 Meaning of Correlation

The term correlation refers to the degree of relationship between two or more variables. If a change in one variable effects a change in the other variable, the variables are said to be correlated.

9.1.2 Types of correlation

Correlation is classified into many types, but the important are

(i) Positive

(ii) Negative

Positive and negative correlation depends upon the direction of change of the variables.

Positive Correlation

If two variables tend to move together in the same direction that is, an increase in the value of one variable is accompanied by an increase in the value of the other variable; or a decrease in the value of one variable is accompanied by a decrease in the value of the other variable, then the correlation is called positive or direct correlation.

Example

(i) The heights and weights of individuals

(ii) Price and Supply

(iii) Rainfall and Yield of crops

(iv) The income and expenditure

Negative Correlation

If two variables tend to move together in opposite direction so that an increase or decrease in the values of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative or inverse correlation.

Example

(i) Price and demand

(ii) Repayment period and EMI

(iii) Yield of crops and price

No Correlation

Two variables are said to be uncorrelated if the change in the value of one variable has no connection with the change in the value of the other variable.

For example

We should expect zero correlation (no correlation) between weight of a person and the colour of his hair or the height of a person and the colour of his hair.

Simple correlation

The correlation between two variables is called simple correlation. The correlation in the case of more than two variables is called multiple correlation.

The following are the mathematical methods of correlation coefficient

(i) Scatter diagram

(ii) Karl Pearson's Coefficient of Correlation

9.1.3 Scatter Diagram

Let $(X_1, Y_1), (X_2, Y_2) \ldots (X_N, Y_N)$ be the N pairs of observation of the variables X and Y. If we plot the values of X along x - axis and the corresponding values of Y along y-axis, the diagram so obtained is called a scatter diagram. It gives us an idea of relationship between X and Y. The type of scatter diagram under a simple linear correlation is given below.

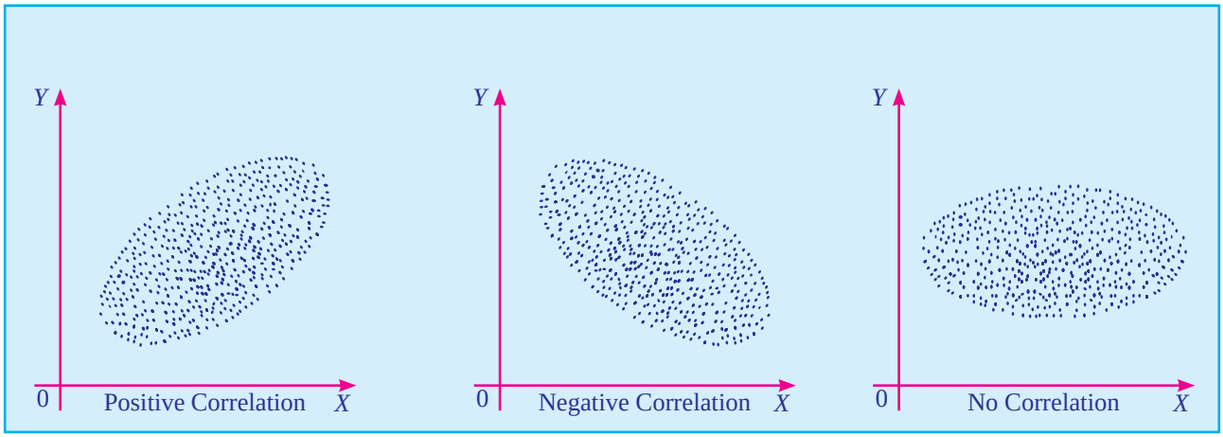


Fig 9.1

(i) If the plotted points show an upward trend, the correlation will be positive.

(ii) If the plotted points show a downward trend, the correlation will be negative.

(iii) If the plotted points show no trend the variables are said to be uncorrelated.

9.1.4 Karl Pearson's Correlation Coefficient

Karl Pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between two variables say X and Y. Karl Pearson's method is the most widely used method in practice and is known as Pearsonian Coefficient of Correlation. It is denoted by the symbol 'r' and defined as

$$r = \frac{\text{cov}(X,Y)}{\overline{\sigma_x}\,\overline{\sigma_y}},$$

where $\text{cov}(X, Y) = \frac{1}{N}\sum(X-\overline{X})(Y-\overline{Y})$

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2}$$

$$\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i-\overline{Y})^2}$$

Hence the formula to compute Karl Pearson Correlation coefficient is

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i-\overline{X})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i-\overline{Y})^2}}$$

$$r = \frac{\sum_{i=1}^{N}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}}$$

Interpretation of Correlation coefficient:

Coefficient of correlation lies between -1 and $+1$. Symbolically, $-1 \le r \le +1$

- When $r = +1$, then there is perfect positive correlation between the variables.
- When $r = -1$, then there is perfect negative correlation between the variables.
- When $r = 0$, then there is no relationship between the variables, that is the variables are uncorrelated.

Thus, the coefficient of correlation describes the magnitude and direction of correlation.

Methods of computing Correlation Coefficient

(i) When deviations are taken from Mean

Of all the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearsonian coefficient of correlation, is most widely used in practice.

$$r = \frac{\sum_{i=1}^{N}(X_i-\overline{X})\,(Y_i-\overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i-\overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Where $x = (X_i-\overline{X})$ and $y = (Y_i-\overline{Y})$;
$i = 1, 2 \ldots N$

This method is to be applied only when the deviations of items are taken from actual means.

Steps to solve the problems:

(i) Find out the mean of the two series that is $\overline{X}$ and $\overline{Y}$.

(ii) Take deviations of the two series from $\overline{X}$ and $\overline{Y}$ respectively and denoted by x and y.

(iii) Square the deviations and get the total of the respective squares of deviation of x and y respectively and it is denoted by Σx^2 and Σy^2.

(iv) Multiply the deviations of x and y and get the total and it is denoted by Σxy.

(v) Substitute the values of Σxy, Σx^2 and Σy^2 in the above formula.

Example 9.1

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	$x = (X-18)$	x^2
6	–12	144
8	–10	100
12	–6	36
15	–3	9
18	0	0
20	2	4
24	6	36
28	10	100
31	13	169
$\sum X = 162$	$\sum x = 0$	$\sum x^2 = 598$

Y	$y = (Y-19)$	y^2	xy
10	–9	81	108
12	–7	49	70
15	–4	16	24
15	–4	16	12
18	–1	1	0
25	6	36	12
22	3	9	18
26	7	49	70
28	9	81	117
$\sum Y = 171$	$\sum y = 0$	$\sum y^2 = 338$	$\sum xy = 431$

Table 9.1

$N=9,\ \overline{X} = \frac{\Sigma X}{N} = \frac{162}{9} = 18, \overline{Y} = \frac{\Sigma Y}{N} = \frac{171}{9} = 19$

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

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

$\Sigma xy = 431,\ \Sigma x^2 = 598,\ \Sigma y^2 = 338$

$$r = \frac{431}{\sqrt{598 \times 338}} = \frac{431}{449.582} = +0.959$$

(ii) When actual values are taken (without deviation)

when the values of X and Y are considerably small in magnitude the following formula can be used

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

Example 9.2

Calculate coefficient of correlation from the following data

X	12	9	8	10	11	13	7
Y	14	8	6	9	11	12	3

Solution:

In both the series items are in small number. Therefore correlation coefficient can also be calculated without taking deviations from actual means or assumed mean.

X	Y	X^2	Y^2	XY
12	14	144	196	168
9	8	81	64	72
8	6	64	36	48
10	9	100	81	90
11	11	121	121	121
13	12	169	144	156
7	3	49	9	21
$\sum X = 70$	$\sum Y = 63$	$\sum X^2 = 728$	$\sum Y^2 = 651$	$\sum XY = 676$

Table 9.2

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

$$= \frac{7(676) - (70)(63)}{\sqrt{7(728) - (70)^2} \times \sqrt{7(651) - (63)^2}}$$

$$= \frac{322}{339.48}$$

$r = +0.95$

(iii) When deviations are taken from an Assumed mean

When actual means are in fractions, say the actual means of X and Y series are 20.167 and 29.23, the calculation of correlation by the method discussed above would involve too many calculations and would take a lot of time. In such cases we make use of the assumed mean method for finding out correlation. When deviations are taken from an assumed mean the following formula is applicable:

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

Where $dx = X - A$ and $dy = Y - B$. Here A and B are assumed mean

NOTE

While applying assumed mean method, any value can be taken as the assumed mean and the answer will be the same. However, the nearer the assumed mean to the actual mean, the lesser will be the calculations.

Steps to solve the problems:

(i) Take the deviations of X series from an assumed mean, denote these deviations by dx and obtain the total that is Σdx.

(ii) Take the deviations of Y series from an assumed mean, denote these deviations by dy and obtain the total that is Σdy.

(iii) Square dx and obtain the total Σdx^2.

(iv) Square dy and obtain the total Σdy^2.

(v) Multiply dx and dy and obtain the total $\Sigma dx\, dy$.

(vi) Substitute the values of $\Sigma\, dxdy$, $\Sigma\, dx$, $\Sigma\, dy$, $\Sigma\, dx^2$ and $\Sigma\, dy^2$ in the formula given above.

Example 9.3

Find out the coefficient of correlation in the following case and interpret.

Height of father (in inches)	65	66	67	67	68	69	71	73
Height of son (in inches)	67	68	64	68	72	70	69	70

Solution:

Let us consider Height of father (in inches) is represented as X and Height of son (in inches) is represented as Y.

X	$dx = (X-67)$	dx^2
65	-2	4
66	-1	1
67	0	0
67	0	0
68	1	1
69	2	4
71	4	16
73	6	36
$\Sigma\ X = 546$	$\Sigma\ dx = 10$	$\Sigma\ dx^2 = 62$

Y	$dy = (Y-68)$	dy^2	$dxdy$
67	–1	1	2
68	0	0	0
64	-4	16	0
68	0	0	0
72	4	16	4
70	2	4	4
69	1	1	4
70	2	4	12
$\Sigma\ Y = 548$	$\Sigma\ dy = 4$	$\Sigma\ dy^2 = 42$	$\Sigma\ dxdy = 26$

Table 9.3

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

Here $\Sigma dx = 10$, $\Sigma dx^2 = 62$, $\Sigma dy = 4$, $\Sigma dy^2 = 42$ and $\Sigma dxdy = 26$

$$r = \frac{(8 \times 26) - (10 \times 4)}{\sqrt{(8 \times 62) - (10)^2} \times \sqrt{(8 \times 42) - (4)^2}}$$

$$r = \frac{168}{\sqrt{396} \times \sqrt{320}}$$

$$r = \frac{168}{355.98} = 0.472$$

i.e. $r = +0.472$

Heights of fathers and their respective sons are positively correlated.

Example 9.4

Calculate the correlation coefficient from the following data

$N=9$, $\sum X = 45$, $\sum Y = 108$, $\Sigma X^2 = 285$, $\Sigma Y^2 = 1356$, $\Sigma XY = 597$.

Solution:

We know that correlation coefficient

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

$$= \frac{9(597) - (45 \times 108)}{\sqrt{9(285) - (45)^2} \times \sqrt{9(1356) - (108^2)}}$$

$r = +0.95$

Example 9.5

From the following data calculate the correlation coefficient $\Sigma xy = 120$, $\Sigma x^2 = 90$, $\Sigma y^2 = 640$.

Solution:

Given $\Sigma xy = 120$, $\Sigma x^2 = 90$, $\Sigma y^2 = 640$

Then $r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{120}{\sqrt{90(640)}} = \frac{120}{\sqrt{57600}}$

$= \frac{120}{240} = 0.5$

9.2 Rank Correlation

9.2.1 Spearman's Rank Correlation Coefficient

In 1904, Charles Edward Spearman, a British psychologist found out the method of ascertaining the coefficient of correlation by ranks. This method is based on rank. This measure is useful in dealing with qualitative characteristics, such as intelligence, beauty, morality, character, etc. It cannot be measured quantitatively, as in the case of Pearson's coefficient of correlation.

Rank correlation is applicable only to individual observations. The result we get from this method is only an approximate one, because under ranking method original value are not taken into account. The formula for Spearman's rank correlation which is denoted by ρ (pronounced as row) is

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

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

where d = The difference of two ranks

$= R_X - R_Y$ and

N = Number of paired observations.

Rank coefficient of correlation value lies between –1 and +1.

Symbolically, $-1 \le \rho \le +1$.

When we come across spearman's rank correlation, we may find two types of problem

(i) When ranks are given

(ii) When ranks are not given

Example 9.6

The following are the ranks obtained by 10 students in Statistics and Mathematics

Statistics	1	2	3	4	5	6	7	8	9	10
Mathematics	1	4	2	5	3	9	7	10	6	8

Find the rank correlation coefficient.

Solution:

Let R_X is considered for the ranks of Statistics and R_Y is considered for the ranks of mathematics.

R_X	R_Y	$d = R_X - R_Y$	d^2
1	1	0	0
2	4	-2	4
3	2	1	1
4	5	-1	1
5	3	2	4
6	9	-3	9
7	7	0	0
8	10	-2	4
9	6	3	9
10	8	2	4
			$\sum d^2 = 36$

Table 9.4

The rank correlation is given by

$$\rho = 1 - \frac{6\Sigma d^2}{N(N^2-1)} = 1 - \frac{6(36)}{10(10^2-1)}$$

$= 1 - 0.218$

$\therefore \rho = 0.782$

Example 9.7

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

First judge	1	4	6	3	2	9	7	8	10	5
Second judge	2	6	5	4	7	10	9	3	8	1
Third judge	3	7	4	5	10	8	9	2	6	1

Use the method of rank correlation coefficient to determine which pair of judges has the nearest approach to common taste in beauty?

Solution:

Let R_X, R_Y, R_Z denote the ranks by First judge, Second judge and third judge respectively

R_X	R_Y	R_Z	$d_{XY} = R_X - R_Y$	$d_{YZ} = R_Y - R_Z$	$d_{ZX} = R_Z - R_X$
1	2	3	-1	-1	2
4	6	7	-2	-1	3
6	5	4	1	1	-2
3	4	5	-1	-1	2
2	7	10	-5	-3	8
9	10	8	-1	2	-1
7	9	9	-2	0	2
8	3	2	5	1	-6
10	8	6	2	2	-4
5	1	1	4	0	-4

d^2_{XY}	d^2_{YZ}	d^2_{ZX}
1	1	4
4	1	9
1	1	4
1	1	4
25	9	64
1	4	1
4	0	4
25	1	36
4	4	16
16	0	16
$\Sigma d^2_{XY} = 82$	$\Sigma d^2_{YZ} = 22$	$\Sigma d^2_{ZX} = 158$

Table 9.5

$$\rho_{XY} = 1 - \frac{6\Sigma d^2_{XY}}{N(N^2-1)} = 1 - \frac{6(82)}{10(10^2-1)}$$

$= 1 - 0.4969 = 0.5031$

$$\rho_{YZ} = 1 - \frac{6\Sigma d^2_{YZ}}{N(N^2-1)} = 1 - \frac{6(22)}{10(10^2-1)}$$

$$= 1-\frac{132}{990} = 1-0.1333=0.8667$$

$$\rho_{ZX} = 1-\frac{6\Sigma d_{ZX}^2}{N(N^2-1)} = 1-\frac{6(158)}{10(10^2-1)}$$

$$= 1-0.9576 = 0.0424$$

Since the rank correlation coefficient between Second and Third judges i.e., ρ_{YZ} is positive and weight among the three coefficients. So, Second judge and Third judge have the nearest approach for common taste in beauty.

Example 9.8

Calculate rank correlation coefficient of the following data

Subject 1	40	46	54	60	70	80	82	85	87	90	95
Subject2	45	46	50	43	40	75	55	72	65	42	70

Solution:

Let X is considered for Subject 1 and

Y is considered for Subject 2.

X	Y	R_X	R_Y	$d = R_X - R_Y$	d^2
40	45	1	4	–3	9
46	46	2	5	–3	9
54	50	3	6	–3	9
60	43	4	3	1	1
70	40	5	1	4	16
80	75	6	11	–5	25
82	55	7	7	0	0
85	72	8	10	–2	4
87	65	9	8	1	1
90	42	10	2	8	64
95	70	11	9	2	4
					$\Sigma d^2 = 142$

Table 9.6

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

$$= 1-\frac{6(142)}{11(11^2-1)}$$

$$= 1-\frac{852}{1320} = 0.354$$

Exercise 9.1

1. Calculate the correlation co-efficient for the following data

X	5	10	5	11	12	4	3	2	7	1
Y	1	6	2	8	5	1	4	6	5	2

2. Find coefficient of correlation for the following:

Cost (₹)	14	19	24	21	26	22	15	20	19
Sales (₹)	31	36	48	37	50	45	33	41	39

3. Calculate coefficient of correlation for the ages of husbands and their respective wives:

Age of husbands	23	27	28	29	30	31	33	35	36	39
Age of wives	18	22	23	24	25	26	28	29	30	32

4. Calculate the coefficient of correlation between X and Y series from the following data

Description	X	Y
Number of pairs of observation	15	15
Arithmetic mean	25	18
Standard deviation	3.01	3.03
Sum of squares of deviation from the arithmetic mean	136	138

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

5. Calculate correlation coefficient for the following data

X	25	18	21	24	27	30	36	39	42	48
Y	26	35	48	28	20	36	25	40	43	39

6. Find coefficient of correlation for the following:

X	78	89	96	69	59	79	68	62
Y	121	72	88	60	81	87	123	92

7. An examination of 11 applicants for a accountant post was taken by a finance company. The marks obtained by the applicants in the reasoning and aptitude tests are given below.

Applicant	A	B	C	D	E	F
Reasoning test	20	50	28	25	70	90
Aptitude test	30	60	50	40	85	90

Applicant	G	H	I	J	K
Reasoning test	76	45	30	19	26
Aptitude test	56	82	42	31	49

Calculate Spearman's rank correlation coefficient from the data given above.

8. The following are the ranks obtained by 10 students in commerce and accountancy are given below:

Commerce	6	4	3	1	2	7	9	8	10	5
Accountancy	4	1	6	7	5	8	10	9	3	2

To what extent is the knowledge of students in the two subjects related?

9. A random sample of recent repair jobs was selected and estimated cost and actual cost were recorded.

Estimated cost	300	450	800	250	500	975	475	400
Actual cost	273	486	734	297	631	872	396	457

Calculate the value of spearman's correlation coefficient.

10. The rank of 10 students of same batch in two subjects *A* and *B* are given below. Calculate the rank correlation coefficient.

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

9.3 Regression Analysis

Introduction:

So far we have studied correlation analysis which measures the direction and strength of the relationship between two variables. Here we can estimate or predict the value of one variable from the given value of the other variable. For instance, price and supply are correlated. We can find out the expected amount of supply for a given price or the required price level for attaining the given amount of supply.

The term " regression" literally means "Stepping back towards the average". It was first used by British biometrician Sir Francis Galton (1822 -1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition 9.1

Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.

9.3.1 Dependent and independent variables

Definition 9.2

In regression analysis there are two types of variables. The variable whose value is to be predicted is called dependent variable and the variable which is used for prediction is called independent variable.

Regression helps us to estimate the value of one variable, provided the value of the other variable is given. The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called Regression.

9.3.2 Regression Equations

Regression equations are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations. The regression equation of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X. Regression equations of (i) X on Y (ii) Y on X and their coefficients in different cases are described as follows.

Case 1: When the actual values are taken

When we deal with actual values of X and Y variables the two regression equations and their respective coefficients are written as follows:

(i) Regression Equation of X on Y:

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

where $\overline{X}$ is the mean of X series,

$\overline{Y}$ is the mean of Y series,

$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$ is known as the regression coefficient of X on Y, and r is the correlation coefficient between X and Y, σ_x and σ_y are standard deviations of X and Y respectively.

(ii) Regression Equation of *Y* on *X*;

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

where $\overline{X}$ is the mean of X series,

$\overline{Y}$ is the mean of Y series,

$b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2}$ is known as the regression coefficient of Y on X, and r is the correlation coefficient between X and Y, σ_x and σ_y are standard deviations of X and Y respectively.

Case 2: Deviations taken from Arithmetic means of X and Y

The calculation can very much be simplified instead of dealing with the actual values of X and Y, we take the deviations of X and Y series from their respective means. In such a case the two regression equations and their respective coefficients are written as follows:

(i) Regression Equation of X on Y:

$$X - \overline{X} = b_{xy}(Y - \overline{Y})$$

where $\overline{X}$ is the mean of X series,

$\overline{Y}$ is the mean of Y series,

$b_{xy} = r\frac{\sigma x}{\sigma y} = \frac{\Sigma xy}{\Sigma y^2}$ is known as the regression coefficient of X on Y, $x = (X - \overline{X})$ and $y = (Y - \overline{Y})$

(ii) Regression Equation of *Y* on *X*;

$Y-\overline{Y} = b_{yx}(X-\overline{X})$

where $\overline{X}$ is the mean of *X* series,

$\overline{Y}$ is the mean of *Y* series,

$b_{yx} = r\frac{\sigma y}{\sigma x} = \frac{\Sigma xy}{\Sigma x^2}$ is known as the regression coefficient of *Y* on *X*, $x = (X-\overline{X})$ and $y = (Y-\overline{Y})$

Note: Instead of finding out the values of correlation coefficient σ_x, σ_y, etc, we can find the value of regression coefficient by calculating Σxy and Σy^2.

Case 3: Deviations taken from Assumed Mean

When actual means of *X* and *Y* variables are in fractions the calculations can be simplified by taking the deviations from the assumed means. The regression equations and their coefficients are written as follows

(i) Regression Equation of Y on X:

$$Y-\overline{Y} = b_{yx}(X-\overline{X})$$

$$b_{yx} = \frac{N\Sigma dxdy-(\Sigma dx)(\Sigma dy)}{N\Sigma dx^2-(\Sigma dx)^2}$$

(ii) Regression Equation of X on Y:

$$X-\overline{X} = b_{xy}(Y-\overline{Y})$$

$$b_{xy} = \frac{N\Sigma dxdy-(\Sigma dx)(\Sigma dy)}{N\Sigma dy^2-(\Sigma dy)^2}$$

where $dx = X-A$, $dy = Y-B$, *A* and *B* are assumed means or arbitrary values are taken from *X* and *Y* respectively.

Properties of Regression Coefficients

(i) Correlation Coefficient is the geometric mean between the regression coefficients $r = \pm\sqrt{b_{xy}\times b_{yx}}$.

(ii) If one of the regression coefficients is greater than unity, the other must be less than unity.

(iii) Both the regression coefficients are of same sign.

Example 9.9

Calculate the regression coefficient and obtain the lines of regression for the following data

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

X	*Y*	X^2	Y^2	X^Y
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
$\Sigma X = 20$	$\Sigma Y = 77$	$\Sigma X^2 = 140$	$\Sigma Y^2 = 875$	$\Sigma XY = 334$

Table 9.7

$\overline{X} = \frac{\Sigma X}{N} = \frac{28}{7} = 4$, $\overline{Y} = \frac{\Sigma Y}{N} = \frac{77}{7} = 11$

Regression coefficient of *X* on *Y*:

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

$$= \frac{7(334)-(28)(77)}{7(875)-(77)^2}$$

$$= \frac{2338-2156}{6125-5929} = \frac{182}{196}$$

$\therefore b_{xy} = 0.929$

(i) Regression equation of *X* on *Y*:

$X-\overline{X} = b_{xy}(Y-\overline{Y})$

$X-4 = 0.929(Y-11)$

$X-4 = 0.929Y-10.219$

$\therefore$ The regression equation *X* on *Y* is $X = 0.929Y-6.219$.

(ii) Regression coefficient *of Y* on *X*:

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

$$= \frac{7(334) - (28)(77)}{7(140) - (28)^2}$$

$$= \frac{2338 - 2156}{980 - 784}$$

$$= \frac{182}{196}$$

$$\therefore b_{yx} = 0.929$$

(iii) Regression equation of *Y* on *X*:

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$$Y - 11 = 0.929\,(X - 4)$$

$$Y = 0.929X - 3.716 + 11$$

$$Y = 0.929X + 7.284$$

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

$$Y = 0.929X + 7.284$$

Example 9.10

Calculate the two regression equations of X on Y and Y on X from the data given below, taking deviations from a actual means of X and Y.

Price (₹)	10	12	13	12	16	15
Amount demanded	40	38	43	45	37	43

Estimate the likely demand when the price is ₹ 20.

Solution:

Calculation of Regression equation

(i) Regression equation of *X* on *Y*:

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

From the Table 9.8, we get

$$\overline{X} = \frac{78}{6} = 13\,,\quad \overline{Y} = \frac{246}{6} = 41$$

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{\Sigma y^2} = \frac{-6}{50} = -0.12$$

$$X - 13 = -0.12\,(Y - 41)$$

$$X - 13 = -0.12Y + 4.92$$

$$X = -0.12Y + 17.92$$

X	$x = (X - 13)$	x^2
10	–3	9
12	–1	1
13	0	0
12	–1	1
16	3	9
15	2	4
$\Sigma X = 78$	$\Sigma x = 0$	$\Sigma x^2 = 24$

Y	$y = (Y - 41)$	y^2	xy
40	–1	1	3
38	–3	9	3
43	2	4	0
45	4	16	–4
37	–4	16	–12
43	2	4	4
$\Sigma Y = 246$	$\Sigma y = 0$	$\Sigma y^2 = 50$	$\Sigma xy = -6$

Table 9.8

(ii) Regression Equation of Y on X:

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

$$b_{yx} = r\,\frac{\sigma_y}{\sigma_x} = \frac{\Sigma xy}{\Sigma x^2} = -\frac{6}{24} = -0.25$$

$$Y - 41 = -0.25\,(X - 13)$$

$$Y - 41 = -0.25\,X + 3.25$$

$$Y = -0.25\,X + 44.25$$

When X is 20, Y will be

$Y = -0.25\,(20)+44.25$

$= -5+44.25$

$= 39.25$ (when the price is ₹20, the likely demand is 39.25)

Example 9.11

Obtain regression equation of Y on X and estimate Y when $X = 55$ from the following:

X	40	50	38	60	65	50	35
Y	38	60	55	70	60	48	30

Solution:

X	Y	$dx = (X-48)$	dx^2	$dy = (Y-50)$	dy^2	$dxdy$
40	38	–8	64	–12	144	96
50	60	2	4	10	100	20
38	55	–10	100	5	25	–50
60	70	12	144	20	400	240
65	60	17	289	10	100	170
50	48	2	4	-2	4	-4
35	30	-13	169	-20	400	260
$\Sigma X =$ 338	$\Sigma Y =$ 361	$\Sigma dx =$ 2	$\Sigma dx^2 =$ 774	$\Sigma dy =$ 11	$\Sigma dy^2 =$ 1173	$\Sigma dxy =$ 732

Table 9.9

$$\overline{X} = \frac{\Sigma X}{N} = \frac{338}{7} = 48.29$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{361}{7} = 51.57$$

(i) Regression coefficients of Y on X:

$$b_{yx} = \frac{N\Sigma dx\,dy - (\Sigma dx)(\Sigma dy)}{N\,\Sigma dx^2 - (\Sigma dx)^2}$$

$$= \frac{7(732)-(2)(11)}{7(774)-(2)^2}$$

$$= \frac{5124-22}{5418-4}$$

$$= \frac{5102}{5414}$$

$= 0.942$

$b_{yx} = 0.942$

(ii) Regression equation of Y on X:

$Y - \overline{Y} = b_{yx}(X - \overline{X})$

$Y-51.57 = 0.942(X-48.29)$

$Y = 0.942X-45.49+51.57=0.942\times -45.49+51.57$

$Y = 0.942X+6.08$

$\therefore$ The regression equation of Y on X is $Y= 0.942X+6.08$

Estimation of Y when $X= 55$

$Y= 0.942(55)+6.08=57.89$

Example 9.12

Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations:

$2Y-X-50 = 0$

$3Y-2X-10 = 0.$

Solution:

We are given

$2Y-X-50 = 0$ ... (1)

$3Y-2X-10 = 0$ ... (2)

Solving equation (1) and (2)

We get $Y= 90$

Putting the value of Y in equation (1)

We get $X = 130$

Hence $\overline{X} = 130$ and $\overline{Y} = 90$

Calculating correlation coefficient

Let us assume equation (1) be the regression equation of Y on X.

$2Y = X+50$

$Y = \frac{1}{2}X+25$ therefore $b_{yx} = \frac{1}{2}$

Clearly equation (2) would be treated as regression equation of X on Y

$3Y-2X-10 = 0$

$2X = 3Y-10$

$X = \frac{3}{2}Y-5$ therefore $b_{xy} = \frac{3}{2}$

The Correlation coefficient $r = \pm\sqrt{b_{xy} \times b_{yx}}$

$r = \sqrt{\frac{1}{2} \times \frac{3}{2}} = 0.866$

NOTE

It may be noted that in the above problem one of the regression coefficient is greater than 1 and the other is less than 1. Therefore our assumption on given equations are correct.

Example 9.13

Find the means of X and Y variables and the coefficient of correlation between them from the following two regression equations:

$4X-5Y+33 = 0$

$20X-9Y-107 = 0$

Solution:

We are given

$4X-5Y+33 = 0$ ... (1)

$20X-9Y-107 = 0$ ... (2)

Solving equation (1) and (2)

We get $Y = 17$

Putting the value of Y in equation (1)

We get $X = 13$

Hence $\overline{X} = 13$ and $\overline{Y} = 17$

Calculating correlation coefficient

Let us assume equation (1) be the regression equation of X on Y

$4X = 5Y-33$

$X = \frac{1}{4}(5Y-33)$

$X = \frac{5}{4}Y-\frac{33}{4}$

$b_{xy} = \frac{5}{4} = 1.25$

Let us assume equation (2) be the regression equation of Y on X

$9Y = 20X-107$

$Y = \frac{1}{9}(20X-107)$

$Y = \frac{20}{9}X-\frac{107}{9}$

$b_{yx} = \frac{20}{9} = 2.22$

But this is not possible because both the regression coefficient are greater than 1. So our above assumption is wrong. Therefore treating equation (1) has regression equation of Y on X and equation (2) has regression equation of X on Y. So we get

$b_{yx} = \frac{4}{5} = 0.8$ and

$b_{xy} = \frac{9}{20} = 0.45$

The Correlation coefficient

$r = \pm\sqrt{b_{xy} \times b_{xy}}$

$r = \sqrt{0.45 \times 0.8} = 0.6$

Example 9.14

The following table shows the sales and advertisement expenditure of a form

Title	Sales	Advertisement expenditure (₹ in Crores)
Mean	40	6
SD	10	1.5

Coefficient of correlation r= 0.9. Estimate the likely sales for a proposed advertisement expenditure of Rs. 10 crores.

Solution:

Let the sales be X and advertisement expenditure be Y

Given $\overline{X}$ =40, $\overline{Y}$ =6, σ_x=10, σ_y=1.5 and r=0.9

Equation of line of regression x on y is

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

$$X-40 = (0.9)\frac{10}{1.5}(Y-6)$$

$$X-40 = 6Y-36$$

$$X = 6Y+4$$

When advertisement expenditure is 10 crores i.e., Y=10 then sales X=6(10)+4=64 which implies sales is 64.

Example 9.15

There are two series of index numbers P for price index and S for stock of the commodity. The mean and standard deviation of P are 100 and 8 and of S are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data obtain the regression lines of P on S and S on P.

Solution:

Let us consider X for price P and Y for stock S. Then the mean and SD for P is considered as $\overline{X}$ =100 and σ_x=8. respectively and the mean and SD of S is considered as $\overline{Y}$ =103 and σ_y = 4. The correlation coefficient between the series is $r(X,Y)$ = 0.4

Let the regression line X on Y be

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

$$X-100 = (0.4)\frac{8}{4}(Y-103)$$

$$X-100 = 0.8(Y-103)$$

X–0.8Y–17.6 = 0 (or) X=0.8Y+17.6

The regression line Y on X be

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

$$Y-103 = (0.4)\frac{4}{8}(X-100)$$

$$Y-103 = 0.2\,(X-100)$$

$$Y-103 = 0.2\,X-20$$

$$Y = 0.2\,X+83 \text{ (or) } 0.2\,X-Y+83=0$$

Example 9.16

For 5 pairs of observations the following results are obtained ΣX=15, ΣY=25, ΣX^2=55, ΣY^2=135, ΣXY=83 Find the equation of the lines of regression and estimate the value of X on the first line when Y=12 and value of Y on the second line if X=8.

Solution:

Here N=5, $\overline{X}=\frac{\Sigma X}{N}=\frac{15}{5}=3$, $\overline{Y}=\frac{\Sigma Y}{N}=\frac{25}{5}=5$ and the regression coefficient

$$b_{xy}=\frac{N\Sigma XY-\Sigma X\Sigma Y}{N\Sigma Y^2-(\Sigma Y)^2}=\frac{5(83)-(15)(25)}{5(135)-(25)^2}=0.8$$

The regression line of X on Y is

$$X-\overline{X} = b_{xy}(Y-\overline{Y})$$

$$X-3 = 0.8(Y-5)$$

$$X = 0.8Y-1$$

When Y = 12, the value of X is estimated as

$$X = 0.8\,(12)-1=8.6$$

The regression coefficient

$$b_{yx} = \frac{N\Sigma XY-\Sigma X\Sigma Y}{N\Sigma X^2-(\Sigma X)^2} = \frac{5(83)-(15)(25)}{5(55)-(15)^2}=0.8$$

Thus b_{yx}= 0.8 then the regression line Y on X is

$$Y-\overline{Y} = b_{yx}(X-\overline{X})$$

$$Y-5 = 0.8(X-3)$$

$$Y = 0.8X+2.6$$

When X = 8 the value of Y is estimated as

$$Y = 0.8(8)+2.6$$

$$Y = 9$$

Example 9.17

The two regression lines are $3X+2Y=26$ and $6X+3Y=31$. Find the correlation coefficient.

Solution:

Let the regression equation of Y on X be

$$3X+2Y = 26$$

$$2Y = -3X+26$$

$$Y = \frac{1}{2}(-3X+26)$$

$$Y = -1.5X+13$$

$$r\frac{\sigma_y}{\sigma_x} = -1.5$$

Implies by $x = r\frac{\sigma_y}{\sigma_x} = -1.5$

Let the regression equation of X on Y be

$$6X+3Y = 31$$

$$X = \frac{1}{6}(-3Y+31) = -0.5Y+5.17$$

$$r\frac{\sigma_x}{\sigma_y} = -0.5$$

Implies $b_{xy} = r\frac{\sigma_x}{\sigma_y} = -0.5$

$$r = \pm\sqrt{b_{xy}.b_{yx}}$$

$= -\sqrt{(-1.5).(-0.5)}$ (Since both the regression coefficient are negative r is negative)

$$\therefore\ r = -0.866$$

Example 9.18

In a laboratory experiment on correlation research study the equation of the two regression lines were found to be $2X-Y+1=0$ and $3X-2Y+7=0$. Find the means of X and Y. Also work out the values of the regression coefficient and correlation between the two variables X and Y.

Solution:

Solving the two regression equations we get mean values of X and Y.

$$2X-Y = -1 \quad \ldots (1)$$

$$3X-2Y = -7 \quad \ldots (2)$$

Solving equation (1) and equation (2) We get $X=5$ and $Y=11$.

Therefore the regression line passing through the means $\overline{X}=5$ and $\overline{Y}=11$.

The regression equation of Y on X is $3X-2Y=-7$

$$2Y = 3X+7$$

$$Y = \frac{1}{2}(3X+7)$$

$$Y = \frac{3}{2}X+\frac{7}{2}$$

$$\therefore\ b_{yx} = \frac{3}{2}(>1)$$

The regression equation of X on Y is

$$2X-Y = -1$$

$$2X = Y-1$$

$$X = \frac{1}{2}(Y-1)$$

$$X = \frac{1}{2}Y-\frac{1}{2}$$

$$\therefore\ b_{xy} = \frac{1}{2}$$

The regression coefficients are positive

$$r = \pm\sqrt{b_{xy}.b_{yx}} = \pm\sqrt{\frac{3}{2}\times\frac{1}{2}}$$

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

$$= \sqrt{\frac{3}{4}}$$

$$= 0.866$$

$$\therefore r = 0.866$$

Example 9.19

For the given lines of regression $3X-2Y=5$ and $X-4Y=7$. Find

(i) Regression coefficients

(ii) Coefficient of correlation

Solution:

(i) First convert the given equations Y on X and X on Y in standard form and find their regression coefficients respectively.

Given regression lines are

$3X-2Y = 5$ ... (1)

$X-4Y = 7$ ... (2)

Let the line of regression of X on Y is

$3X-2Y = 5$

$3X = 2Y+5$

$X = \frac{1}{3}(2Y+5)$

$X = \frac{1}{3}(2Y+5)$

$X = \frac{2}{3}Y+\frac{5}{3}$

$\therefore$ Regression coefficient of X on Y is

$b_{xy} = \frac{2}{3}(<1)$

Let the line of regression of Y on X is

$X-4Y = 7$

$-4Y = -X+7$

$4Y = X-7$

$Y = \frac{1}{4}(X-7)$

$Y = \frac{1}{4}X-\frac{7}{4}$

$\therefore$ Regression coefficient of Y on X is

$b_{yx} = \frac{1}{4}(<1)$

(ii) **Coefficient of correlation**

Since the two regression coefficients are positive then the correlation coefficient is also positive and it is given by

$r = \sqrt{b_{yx} \cdot b_{xy}}$

$= \sqrt{\frac{2}{3} \cdot \frac{1}{4}}$

$= \sqrt{\frac{1}{6}}$

$= 0.4082$

$\therefore r = 0.4082$

Exercise 9.2

1. From the data given below:

Marks in Economics:	25	28	35	32	31
Marks in Statistics:	43	46	49	41	36
Marks in Economics:	36	29	38	34	32
Marks in Statistics:	32	31	30	33	39

Find (a) The two regression equations, (b) The coefficient of correlation between marks in Economics and statistics, (c) The mostly likely marks in Statistics when the marks in Economics is 30.

2. The heights (in cm.) of a group of fathers and sons are given below:

Heights of fathers:	158	166	163	165	167	170	167	172	177	181
Heights of Sons :	163	158	167	170	160	180	170	175	172	175

Find the lines of regression and estimate the height of son when the height of the father is 164 cm.

3. The following data give the height in inches (X) and the weight in lb. (Y) of a random sample of 10 students from a large group of students of age 17 years:

X	61	68	68	64	65	70	63	62	64	67
Y	112	123	130	115	110	125	100	113	116	125

Estimate weight of the student of a height 69 inches.

4. Obtain the two regression lines from the following data $N=20$, $\Sigma X=80$, $\Sigma Y=40$, $\Sigma X^2=1680$, $\Sigma Y^2=320$ and $\Sigma XY=480$.

5. Given the following data, what will be the possible yield when the rainfall is 29.

Details	Rainfall	Production
Mean	25``	40 units per acre
Standard Deviation	3``	6 units per acre

Coefficient of correlation between rainfall and production is 0.8.

6. The following data relate to advertisement expenditure (in lakh of rupees) and their corresponding sales (in crores of rupees)

Advertisement expenditure	40	50	38	60	65	50	35
Sales	38	60	55	70	60	48	30

Estimate the sales corresponding to advertising expenditure of ₹ 30 lakh.

7. You are given the following data:

Details	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8

If the Correlation coefficient between X and Y is 0.66, then find (i) the two regression coefficients, (ii) the most likely value of Y when $X = 10$.

8. Find the equation of the regression line of Y on X, if the observations (X_i, Y_i) are the following (1,4) (2,8) (3,2) (4,12) (5, 10) (6, 14) (7, 16) (8, 6) (9, 18).

9. A survey was conducted to study the relationship between expenditure on accommodation (X) and expenditure on Food and Entertainment (Y) and the following results were obtained:

Details	Mean	SD
Expenditure on Accommodation (₹)	178	63.15
Expenditure on Food and Entertainment (₹)	47.8	22.98
Coefficient of Correlation	0.43	

Write down the regression equation and estimate the expenditure on Food and Entertainment, if the expenditure on accommodation is ₹ 200.

10. For 5 observations of pairs of (X, Y) of variables X and Y the following results are obtained. $\Sigma X = 15$, $\Sigma Y = 25$, $\Sigma X^2 = 55$, $\Sigma Y^2 = 135$, $\Sigma XY = 83$. Find the equation of the lines of regression and estimate the values of X and Y if $Y = 8$; $X = 12$.

11. The two regression lines were found to be $4X-5Y+33 = 0$ and $20X-9Y-107 = 0$. Find the mean values and coefficient of correlation between X and Y.

12. The equations of two lines of regression obtained in a correlation analysis are the following $2X = 8-3Y$ and $2Y = 5-X$. Obtain the value of the regression coefficients and correlation coefficient.

Exercise 9.3

Choose the correct answer

1. Example for positive correlation is
   (a) Income and expenditure
   (b) Price and demand
   (c) Repayment period and EMI
   (d) Weight and Income

2. If the values of two variables move in same direction then the correlation is said to be
   (a) Negative
   (b) Positive
   (c) Perfect positive
   (d) No correlation

IL8BC4

3. If the values of two variables move in opposite direction then the correlation is said to be

(a) Negative
(b) Positive
(c) Perfect positive
(d) No correlation

4. Correlation co-efficient lies between
(a) 0 to ∞ (b) –1 to +1
(c) –1 to 0 (d) –1 to ∞

5. If $r(X,Y) = 0$ the variables X and Y are said to be
(a) Positive correlation
(b) Negative correlation
(c) No correlation
(d) Perfect positive correlation

6. The correlation coefficient from the following data N=25, ΣX=125, ΣY=100, ΣX^2=650, ΣY^2=436, ΣXY=520
(a) 0.667 (b) -0.006
(c) –0.667 (d) 0.70

7. From the following data, N=11, ΣX=117, ΣY=260, ΣX^2=1313, ΣY^2=6580, ΣXY=2827the correlation coefficient is
(a) 0.3566 (b) –0.3566
(c) 0 (d) 0.4566

8. The correlation coefficient is
(a) $r(X,Y)=\frac{\sigma_x \sigma_y}{\text{cov}(x,y)}$
(b) $r(X,Y)=\frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$
(c) $r(X,Y)=\frac{\text{cov}(x,y)}{\sigma_y}$
(d) $r(X,Y)=\frac{\text{cov}(x,y)}{\sigma_x}$

9. The variable whose value is influenced (or) is to be predicted is called
(a) dependent variable
(b) independent variable
(c) regressor
(d) explanatory variable

10. The variable which influences the values or is used for prediction is called
(a) Dependent variable
(b) Independent variable
(c) Explained variable
(d) Regressed

11. The correlation coefficient
(a) $r=\pm\sqrt{b_{xy}\times b_{yx}}$
(b) $r=\frac{1}{b_{xy}\times b_{yx}}$
(c) $r=b_{xy}\times b_{yx}$
(d) $r=\pm\sqrt{\frac{1}{b_{xy}\times b_{yx}}}$

12. The regression coefficient of X on Y
(a) $b_{xy}=\frac{N\Sigma dx\,dy-(\Sigma dx)(\Sigma dy)}{N\,\Sigma dy^2-(\Sigma dy)^2}$
(b) $b_{yx}=\frac{N\Sigma\,dx\,dy-(\Sigma dx)(\Sigma dy)}{N\,\Sigma dy^2-(\Sigma dy)^2}$
(c) $b_{xy}=\frac{N\Sigma dxdy-(\Sigma dx)(\Sigma dy)}{N\,\Sigma dx^2-(\Sigma dx)^2}$
(d) $b_{y}=\frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{N\Sigma x^2-(\Sigma x)^2}\times\sqrt{N\Sigma y^2-(\Sigma y)^2}}$

13. The regression coefficient of Y on X
(a) $b_{xy}=\frac{N\Sigma dx\,dy-(\Sigma dx)(\Sigma dy)}{N\,\Sigma dy^2-(\Sigma dy)^2}$
(b) $b_{yx}=\frac{N\Sigma dxdy-(\Sigma dx)(\Sigma dy)}{N\Sigma dy^2-(\Sigma dy)^2}$
(c) $b_{yx}=\frac{N\Sigma dx\,dy-(\Sigma dx)(\Sigma dy)}{N\Sigma dx^2-(\Sigma dx)^2}$
(d) $b_{xy}=\frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{N\Sigma x^2-(\Sigma x)^2}\times\sqrt{N\Sigma y^2-(\Sigma y)^2}}$

14. When one regression coefficient is negative, the other would be
(a) Negative (b) Positive
(c) Zero (d) None of them

15. If X and Y are two variates, there can be atmost
    (a) One regression line
    (b) two regression lines
    (c) three regression lines
    (d) more regression lines

16. The lines of regression of X on Y estimates
    (a) X for a given value of Y
    (b) Y for a given value of X
    (c) X from Y and Y from X
    (d) none of these

17. Scatter diagram of the variate values (X,Y) give the idea about
    (a) functional relationship
    (b) regression model
    (c) distribution of errors
    (d) no relation

18. If regression co-efficient of Y on X is 2, then the regression co-efficient of X on Y is
    (a) $\leq \frac{1}{2}$ (b) 2
    (c) $> \frac{1}{2}$ (d) 1

19. If two variables moves in decreasing direction then the correlation is
    (a) positive (b) negative
    (c) perfect negative (d) no correlation

20. The person suggested a mathematical method for measuring the magnitude of linear relationship between two variables say X and Y is
    (a) Karl Pearson
    (b) Spearman
    (c) Croxton and Cowden
    (d) Ya Lun Chou

21. The lines of regression intersect at the point
    (a) (X,Y) (b) $(\overline{X}, \overline{Y})$
    (c) $(0,0)$ (d) (σ_x, σ_y)

22. The term regression was introduced by
    (a) R.A. Fisher
    (b) Sir Francis Galton
    (c) Karl Pearson
    (d) Croxton and Cowden

23. If $r=-1$, then correlation between the variables
    (a) perfect positive
    (b) perfect negative
    (c) negative
    (d) no correlation

24. The coefficient of correlation describes
    (a) the magnitude and direction
    (b) only magnitude
    (c) only direction
    (d) no magnitude and no direction

25. If $\text{Cov}(x,y)=-16.5$, $\sigma_x^2=2.89$, $\sigma_y^2=100$. Find correlation coefficient.
    (a) –0.12 (b) 0.001
    (c) –1 (d –0.97

Miscellaneous Problems

1. Find the coefficient of correlation for the following data:

X	35	40	60	79	83	95
Y	17	28	30	32	38	49

2. Calculate the coefficient of correlation from the following data:

   $\Sigma X=50$, $\Sigma Y=-30$, $\Sigma X^2=290$, $\Sigma Y^2=300$, $\Sigma XY=-115$, $N=10$

3. Calculate the correlation coefficient from the data given below:

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

4. Calculate the correlation coefficient from the following data:
$\Sigma X=125$, $\Sigma Y=100$, $\Sigma X^2=650$, $\Sigma Y^2=436$, $\Sigma XY=520$, $N=25$

5. A random sample of recent repair jobs was selected and estimated cost , actual cost were recorded.

Estimated cost	30	45	80	25	50	97	47	40
Actual cost	27	48	73	29	63	87	39	45

Calculate the value of spearman's correlation.

6. The following data pertains to the marks in subjects A and B in a certain examination. Mean marks in A = 39.5, Mean marks in B= 47.5 standard deviation of marks in A =10.8 and Standard deviation of marks in B = 16.8. coefficient of correlation between marks in A and marks in B is 0.42. Give the estimate of marks in B for candidate who secured 52 marks in A.

7. X and Y are a pair of correlated variables. Ten observations of their values (X,Y) have the following results. $\Sigma X=55$, $\Sigma XY=350$, $\Sigma X^2=385$, $\Sigma Y=55$, Predict the value of y when the value of X is 6.

8. Find the line regression of Y on X

X	1	2	3	4	5	8	10
Y	9	8	10	12	14	16	15

9. Using the following information you are requested to (i) obtain the linear regression of Y on X (ii) Estimate the level of defective parts delivered when inspection expenditure amounts to ₹ 82 $\Sigma X=424$, $\Sigma Y=363$, $\Sigma X^2=21926$, $\Sigma Y^2=15123$, $\Sigma XY=12815$, $N=10$. Here X is the expenditure on inspection, Y is the defective parts delivered.

10. The following information is given.

Details	X (in ₹)	Y (in ₹)
Arithmetic Mean	6	8
Standard Deviation	5	$\frac{40}{3}$

Coefficient of correlation between X and Y is $\frac{8}{15}$. Find (i) The regression Coefficient of Y on X (ii) The most likely value of Y when X = ₹ 100.

Case Study-1

Mr. Bean visited a departmental store in Triplicane at Chennai on 1st March 2018 and chooses 15 different types of food items that include nutrition information on its packaging. For each food Mr. Bean observed and identified the amount of fat (gms) and the sodium content (mgs/100gms) per serving and recorded in the following table.

Sl. No.	Product Items	Fat (gm/100gms)	Sodium (mg / 100gms)
1.	Dates	0.4	74.4
2.	Appalam	0.26	1440
3.	Energy drink	1.8	136
4.	Gulabjamun Powder	10.4	710
5.	Atta	2.2	4.97
6.	Athi fruits	0.14	2
7.	Alu Muttar mix	5	440

8.	Popcorn	2.32	51.38
9.	Perungayum	0.37	40
10.	Mushroom	31	11.73
11.	Friut juice	0.1	74
12.	Choclate	0.8	0.09
13.	Rava	9	575
14.	Biscuit	19.7	498
15.	Snacks	33.5	821

Mr. Bean wants to establish some statistical relationship between the above mentioned food contents. Here the variable under study are X and Y which represents the amount of Fat content and the amount of Sodium content in each food items respectively. Thus Mr. Bean gets a pair of values (X,Y) for each food item. Mr. Bean further found the average fat content in all the 15 food items is $\overline{X}$ =7.8 (gms/100g) and the average sodium content in all the items is $\overline{Y}$ =325.23(mg/100g). Further, it was identified that the minimum amount of fat contained in Tropicana fruit juice is 0.1(g/100gms) and the maximum amount of fat contained in Lays is 33.5(g/100gms). Thus the fat contained in all the 15 food items is ranging from 0.1(g/100gms) to 33.5(g/100gms). Similarly, the minimum amount of sodium content contained in Kelloggys Choco is 0.09(mg/100gm) and the maximum amount of sodium content in Bhindhu appalam is 1440(mg/100gm). So, the measure of range of fat content and sodium content in all the 15 items are 33.4gm and 1439.91(mg/100gm) respectively. Besides, Mr. Bean is interested in knowing the variation of each individual item from the mean of all observations. He attempted another measure of dispersion known as standard deviation. The measure of standard deviation indicates that there is an average deviation of 11.3 (g/100gms) in fat content and 420.14(mg/100gms) in sodium content of all the 15 food items. Further, Mr. Bean is keen on finding the association between the variables X and Y. So, the correlation analysis has been carried out. The correlation coefficient $r(X,Y)$=0.2285 indicates that there is 22.86% positive association between sodium content and the amount of fat content. Thus from this study Mr. Bean inferred and convinced that the nutrition information on the packaging of each food item is sufficient.

Case Study-2

We collected data relating to the gold price (per gram) in two places namely Chennai Market and Mumbai Market for ten days from 20th Feb 2018 to 1st March 2018 and the same is recorded below.

Date	20th Feb	21st Feb	22nd Feb	23rd Feb	24th Feb
Chennai X	2927	2912	2919	2912	2921
Mumbai Y	2923	2910	2907	2920	2919

Date	25th Feb	26th Feb	27th Feb	28th Feb	1st Mar
Chennai X	2921	2927	2924	2908	2893
Mumbai Y	2919	2925	2918	2902	2895

Do we agree that the price of gold in Chennai market has its impact on Mumbai market?

Let X denotes the gold price per gm in Chennai market and Y denotes the gold

price per gm in Mumbai market. The actual observations given in the table indicates that the gold price ranges from ₹ 2893 to ₹ 2927 in Chennai market and the gold price range from ₹ 2895 (per gram) to ₹ 2925 (per gram) in Mumbai market. Further it is to be noted that there is some oscillations in the gold price rate dated between 20th Feb to 24th Feb and 25th Feb price is remain same as the previous day that is Feb 24th. It may be due to holiday of Gold markets. It is clear from the above table the price of gold rate go on rapidly decreasing from 27th Feb to 1st March. The same fluctuations is observed in Mumbai market from Feb 20th to 24th and remain same on 24th and 25th and rapidly decreasing from Feb 27th to 1st March . So, the market trend in respect of gold rate is same in two markets. We found that the average gold price in Chennai market during these 10 days is ₹ 2916.4 (per gm) and the Mumbai market is ₹ 2913.8 (per gm). The variation among the prices of gold in 10 days. To identify the variation of each individual observation from the mean of all observations. We make use of the best measure known as standard deviation. In this study it is found that the price of gold has a average deviation of ₹ 10 (approximately) in Chennai market and ₹ 9 (approximately) in Mumbai market. To verify the consistency of the prices between the two cities, we attempt coefficient of variation which expresses the percentage of variation. Coefficient of variation of gold price in Chennai market is

$$CV_X = \frac{\sigma_X}{\overline{X}} \times 100 = \frac{10}{2916.4} \times 100 = 0.343\%$$

Similarly, Coefficient of variation of gold price in Mumbai market is

$$CV_Y = \frac{\sigma_Y}{\overline{Y}} \times 100 = \frac{9}{2913.8} \times 100 = 0.31\%$$

Comparison of these coefficients of variations we inferred that the Mumbai market has consistent or more stable in price of gold.

Further to examine the linear relationship between the two variables X and Y. We carry out correlation analysis results $r(X,Y) = 0.8682$. It indicates that there is a positive correlation in price of gold between Mumbai market and Chennai market. Do you think finding a regression line makes sense here?

Summary

- The term correlation refers to the degree of relationship between two or more variables.
- Scatter diagram is a graphic device for finding correlation between two variables.
- Karl Pearson correlation coefficient $r(x,y)=\frac{\text{cov}(X,Y)}{\sigma_x \sigma_y}$
- Correlation coefficient r lies between -1 and 1. (i.e) $-1 \leq r \leq 1$
- When $r=+1$, then the correlation is perfect positive
- When $r=-1$, then the correlation is perfect negative
- When $r=0$, then there is no relationship between the variables, (i.e) the variables are uncorrelated.
- Rank correlation deals with qualitative characteristics.
- Spearman's rank correlation coefficient formula ρ is given by

  $\rho = 1-\frac{6\Sigma d^2}{N(N^2-1)}$ where d = The difference between two ranks = $R_X - R_Y$ and

  N = Number of paired observations.
- Correlation represents linear relationship between the variables but the regression helps to estimate (or predict) one variable by using the other variable.
- Regression lines of

  (i) X on Y is $X-\overline{X}=b_{xy}(Y-\overline{Y})$ (ii) Y on X is $Y-\overline{Y}=b_{yx}(X-\overline{X})$
- The two regression lines passing through their respective means of X and Y
- Calculation of the regression coefficients.

  $b_{xy}=r\frac{\sigma_x}{\sigma_y}$ and $b_{yx}=r\frac{\sigma_y}{\sigma_x}$
- The Properties of regression coefficients.

  (i) $r=\sqrt{b_{yx}\times b_{xy}}$

  (ii) both the regression coefficients cannot be greater than one.

  (iii) Both the regression coefficients have same sign.

GLOSSARY (கலைச்சொற்கள்)	
Abnormal	அசாதாரனமான
Approximate	தோராயமாக
Assumed Mean	ஊகிக்கப்பட்ட சராசரி
Bivariate analysis	இருமாறி பகுப்பாய்வு
Characteristics	பண்புகள்
Closeness	பொருத்தமுடைய
Corrleation	ஒட்டுறவு
Deviations	விலக்கம்
Fluctuate	ஏற்ற இறக்கம்
Interpretation	விளக்கம்
Negative Correlation	எதிர்மறை ஒட்டுறவு
Positive Correlation	நேரிடை ஒட்டுறவு
Random variables	சமவாய்ப்பு மாறிகள்
Regression analysis	தொடர்பு போக்கு ஆய்வு
Relative Variable	சார்ந்த மாறி
Univariate analysis	ஒருமாறி பகுப்பாய்வு

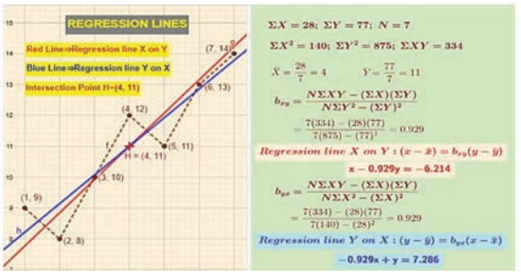
ICT Corner

Correlation and regression analysis

Step – 1

Open the Browser type the URL Link given below (or) Scan the QR Code.

GeoGebra Workbook called "**11th BUSINESS MATHEMATICS and STATISTICS**" will appear. In that there are several worksheets related to your Text Book.



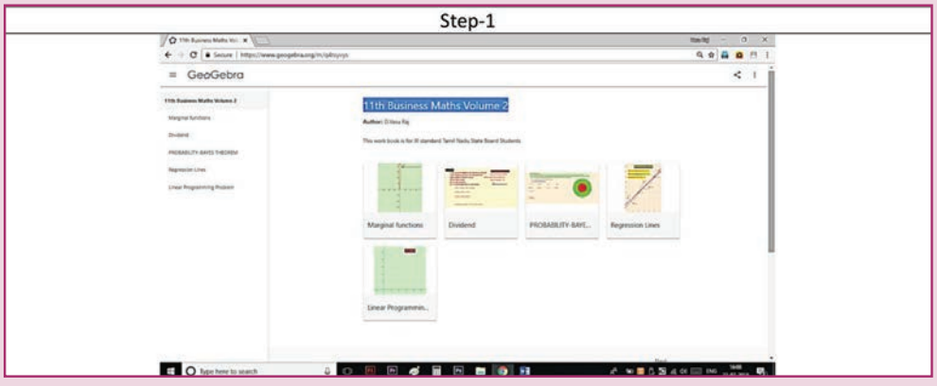
Expected Outcome ⇒

Step-2

Select the work sheet "Regression lines" work out the problem for the data given and workout as given and verify the steps. See the graph of regression line x on y and regression line y on x and check the intersection of these two lines. (Mean of x, mean of y) you can change the data x and Y in the spread sheet for new problem.

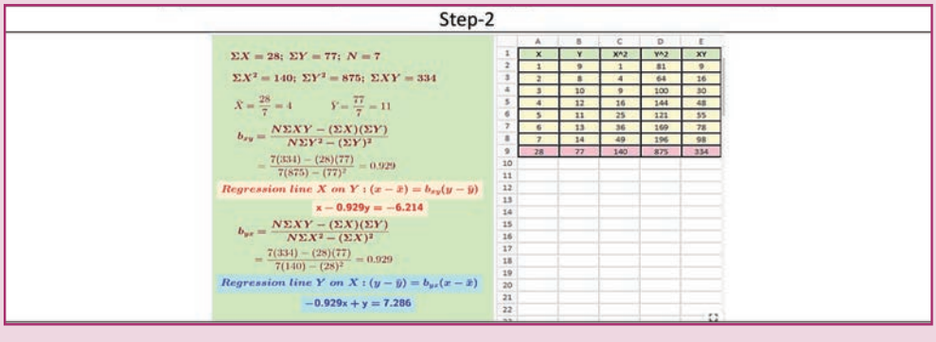
Step 1

Step-1



Step 2

Step-2



Browse in the link

11th Business Mathematics and Statistics
https://ggbm.at/qKj9gSTG (or) scan the QR Code

B184_11_BM_EM