3 CHAPTER

CORRELATION AND REGRESSION ANALYSIS

CHAPTER OUTLINE

After studying this chapter, the reader will be able to understand the

- Correlation: Definition, Scatter diagram, Karl Pearson's Coefficient of Correlation, Numerical problems for determination of Correlation Coefficients.
- Regression: Definition, Dependent and Independent Variables, Least Square method only, Numerical Problems.

BIVARIATE DATA

When data are collected according to a single variable is called univariate data. For example, data on age of individuals. When data is collected according to two variables then it is called bivariate data. For example, data on height and weight of students.

Bivariate Frequency Distribution

Frequency distribution involving only one variable is called univariate frequency distribution. In all situation univariate distribution is not sufficient to represent data. In many situations simultaneous study of two variables become necessary. For example, we want to classify data relating to the weight and height of a group of individuals, income and expenditure of a group of individuals, age of husbands and wives etc.

The data so classified on the basis of two variables give rise to the so called bivariate frequency distribution and it can be summarized in the form of a table is called bivariate (two-way) frequency table. While preparing a bivariate frequency distribution, the values of each variable are grouped into various classes (not necessarily the same for each variable). If the data corresponding to one variable, say X is grouped into m classes and the data corresponding to the other variable, say Y is grouped into n classes then the bivariate table will consist of mxn cells. By going through the different pairs of the values, (X,Y) of the variables and using tally marks we can find the frequency of each cell and thus, obtain the bivariate frequency distribution. The format of a bivariate frequency distribution is given below:

Format of Bivariate Frequency distribution

<table>
<tr><td colspan="2" rowspan="2">X-series
Y-series</td><td>Class -Intervals
Mid Values</td><td rowspan="2">Marginal Frequency of Y</td></tr>
<tr><td>Class - Intervals</td><td>Mid Values</td><td>frequency of the pair
f(x,y)</td></tr>
<tr><td colspan="2">Marginal Frequency of X</td><td>fx</td><td>Total $\Sigma fx = \Sigma fy = N$</td></tr>
</table>

Here f(x, y) is the frequency of the pair (x, y). The frequency distribution of the values of the variables x together with their frequency total (fx) is called the marginal distribution of x and the frequency distribution of the values of the variable Y together with the total frequencies is known as the marginal frequency distribution of Y. The total of the values of manual frequencies is called grand total (N).

For example: the data given below relate to the height and weight of 20 persons. Construct a bivariate frequency table with class interval of height as 62-64,64-66 ... andweightas115-125,13...

S.No.	Height	Weight	S.No.	Height	Weight
1	70	170	11	70	163
2	65	135	12	67	139
3	65	136	13	63	122
4	64	137	14	68	134
5	69	148	15	67	140
6	63	121	16	69	132
7	65	117	17	65	120
8	70	128	18	68	148
9	71	143	19	67	129
10	62	129	20	67	152

Bivariate frequency table showing height and weight of persons.

Height(X) / Weight(Y)	62-64	64-66	66-68	68-70	70-72	Total
115-125	II (2)	II (2)				4
125-135	I	(1)	I (1)	II (2)	I (1)	5
135-145		III (3)	II (2)		I (1)	6
145-155			I (1)	II (2)		3
155-165					I (1)	1
165-175					I (1)	1
Total	3	5	4	4	4	20

CORRELATION

The term correlation is used by a common man without knowing the meaning of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

The study related to the characteristics of only one variable such as height, weight, age, marks, wage etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-variate Analysis. Sometimes the variables may be inter-related. We study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure etc. The nature and strength of relationship may be examined by correlation.

Thus correlation refers to the relationship of two variables or more. For example, relation between height of father and son, yield of crop and rainfall, wage and price index etc. Correlation is statistical Analysis which measures and analyze the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. If two or more variables are so related that the change of one variable brings re change in the value of other variable, then the variables are said to be correlated. This relationship between the variables is called Correlation. Hence two variables are said to be correlated if change in one variable is accompanied by change in other variable.

For Example: The change in quantity of irrigation brings the changes in production. So irrigation and production are correlated. The measure of Correlation is called the Correlation. It measures the degree and direction of relationship between the variables but not say which is cause and which is effect.

Types of Correlation

It can be classified by following three types

1. Positive and Negative Correlation
2. Linear and Non-Linear Correlation
3. Simple, Multiple and Partial Correlation

1. Positive and Negative Correlation

If both the variables move in the same direction i.e. if the increase or decrease in the value of one variable results the increase or decrease in the value of other variable, then the two variables are said to Positive Correlated. For example:

Age(year)	5	8	10	13	15
Weight(kg)	20	28	32	40	45

Similarly, if both variables move in opposite direction, then correlation between the two variables is called Negative Correlation. In this correlation, if the value of one increases then the value of other variable decrease & vice versa. For example:

x	15	20	30	40	50
y	25	20	10	8	2

The relationship between two variables can be measured either by graphical method (Scatter diagram) or by numerical calculated method.

2. Linear and Non-linear Correlation:

The correlation between two variables is said to be linear if corresponding to a unit change in the other variable over the entire range of the value

For example,

X	6	7	8	9	10
Y	5	7	9	11	13

In the other hand, if corresponding to a unit change in one variable, there is no constant change in other variable then the correlation is said to be Non-linear.

X	1	2	3	4	5
Y	7	8	15	20	23

3. Simple, Multiple and partial correlation

The relationship between two (i.e. one dependent and one independent) variables is called Simple Correlation.

The study of relationship among three or more variables simultaneously (at the same time) is known as multiple correlations. In this correlation all the given variables are studied at one time by taking one variable as dependent and the entire remaining variable as independent and the study of relationship between two variables keeping the effect of all remaining variables

Simple Correlation

It is the relationship between two variables. Two variables are correlated if change in one variable is accompanied by change in other variable.

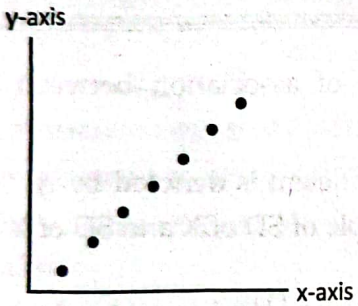
These are the various method of various method of studying the correlation between two variables:

(a) Scatter Diagram Method
(b) Karl Pearson's Correlation coefficient
(c) Spearman's Rank Correlation

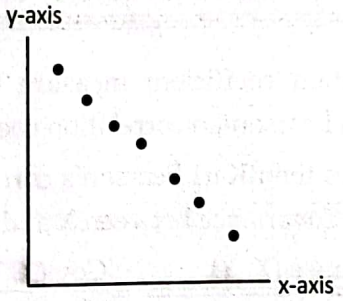
Scatter Diagram Method

Scatter diagram is one of the simplest ways of diagrammatic representation of bi-variate distribution and provides us one of simplest and attractive tools of ascertaining the correlation between two variables. In this method, the points are represented by dots by keeping the independent variable on x-axis and the dependent variable on y-axis. The graph formed by plotting these pairs of x and y is known as Scatter diagram. If the plotted dots show upward or downward then two variables under study are correlated. If the dots are close together and follows some trend of either increasing or decreasing then there is a strong relationship between them.

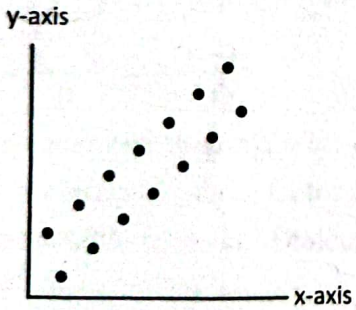
The following diagram of the scattered data depict different forms of correlation:



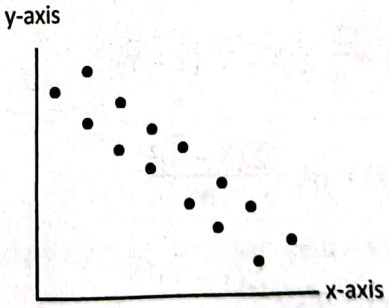
Perfect Positive Correlation



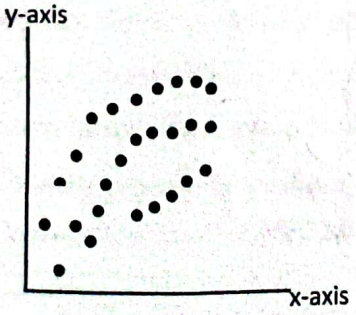
Perfect Negative correlation



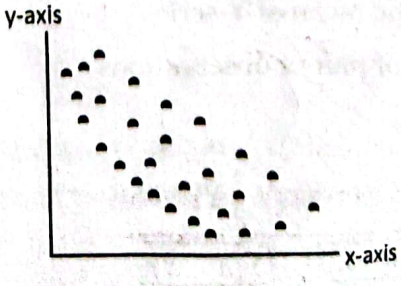
High Degree Positive Correlation



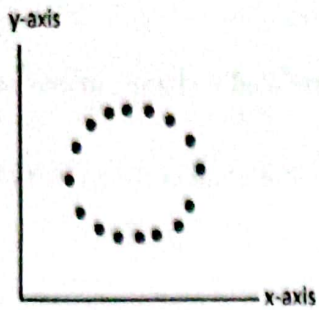
High Degree Negative Correlation



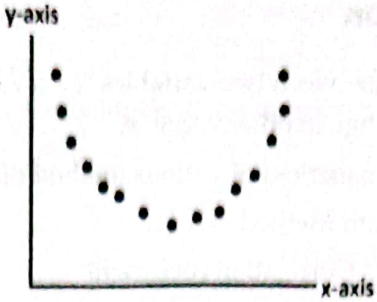
Low degree of positive correlation



Low degree of negative correlation



No correlation



No correlation

Merits

a. It is the simplest method of measuring correlation.

b. It is least affected by the size of extra values.

c. It can be easily understood by non-statistician person.

Demerits

It gives the only rough idea. So it cannot give the extra idea about the correlation between two variables.

KARL PEARSON'S CORRELATION COEFFICIENT

The Karl Pearson's correlation coefficient measure the degree of association between the two variables. It is also known as Pearsonian correlation coefficient.

Let X and Y be two variables then Karl Pearson's correlation coefficient is denoted by r_{xt} or r_{yx} or simply r is define as ratio of Covariance between X and Y to multiple of SD of X and SD of Y

$$r = \frac{\text{Covriance (X, Y)}}{\sqrt{\text{Variance X}}\sqrt{\text{Variance (Y)}}} = \frac{\text{Cov (X, Y)}}{\sigma_X \sigma_Y} \qquad \ldots\ldots (1)$$

Where

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

$$\text{Variance (x)} = \sigma^2_x = \frac{\Sigma(X - \bar{X})^2}{n}$$

$$\text{Variance (y)} = \sigma^2_y = \frac{\Sigma(Y - \bar{Y})^2}{n}$$

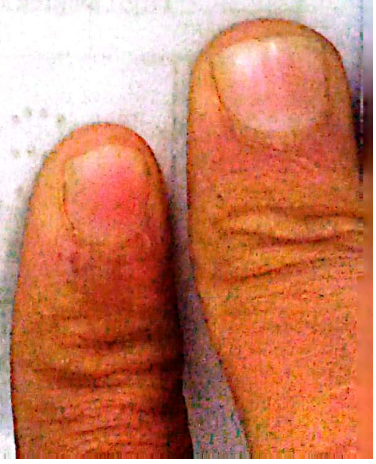
$\bar{X}$ = Arithmetic mean of X-series

$\bar{Y}$ = Arithmetic mean of Y-series

n = Number of pair of observations

Now

$$r = \frac{\frac{1}{n}\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{\Sigma(X - \bar{X})^2}{n}}\sqrt{\frac{\Sigma(Y - \bar{Y})^2}{n}}}$$



$$= \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2}\sqrt{\sum(Y - \bar{Y})^2}}$$

$$= \frac{\sum xy}{\sqrt{(\sum x^2)}\sqrt{\sum y^2}} \quad \ldots\ldots (2)$$

Where, $x = X - \bar{X}$, $Y = Y - \bar{Y}$

Formula (2) is known as product moment method.

After simplification of (1)

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\frac{\sum(X - \bar{X})(Y - \bar{Y})}{n}}{\sqrt{\frac{\sum(X - \bar{X})}{n}}\sqrt{\frac{\sum(Y - \bar{Y})}{n}}} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

Direct Method

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} \quad \ldots\ldots (3)$$

Deviation Method (Change of Origin)

$$r = \frac{n\sum uv - \sum u \sum v}{\sqrt{n\sum u^2 - (\sum u)^2}\sqrt{n\sum v^2 - (\sum v)^2}}$$

Where $u = x - A$; $v = y - B$

A = assumed mean of X-series

B = assumed mean of Y-series

This is the suitable method of calculating the correlation coefficient of two variables consisting large numerical size.

Step Deviation Method (Change of Origin and Scale)

$$r = \frac{n\sum u'v' - \sum u' \sum v'}{\sqrt{n\sum u'^2 - (\sum u')^2}\sqrt{n\sum v'^2 - (\sum v')^2}}$$

$$u' = \frac{x - A}{h}, \; v' = \frac{y - B}{k}$$

Where h = common (scale) factor of X-series

k = common (scale) factor of Y-series

This is the suitable method of calculating the correlation having common factor in x-series or y-series or both series.

Properties of Correlation Coefficient

Following are the important properties of Karl Pearson's Correlation Coefficient

i. Correlation coefficient lies between -1 to 1 i.e., $-1 \leq r \leq +1$.

ii. Correlation coefficient is symmetrical i.e. $r_{xy} = r_{yx} = r$.

iii. Correlation coefficient is independent of change of origin and scale.

iv. Correlation coefficient is the geometric mean of two regression coefficient

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

v. Correlation coefficient has no unit because of relative measure.

Interpretation of calculated value of r

i. If r = + 1, there is Perfect Positive Correlation between two variables.

ii. If r = -1, there is Perfect Negative Correlation between two variables

iii. If r = 0, there is no correlation between two variables or uncorrelated between two variable

iv. If r lies between 0.001 to 0.499 there is low degree of positive correlation

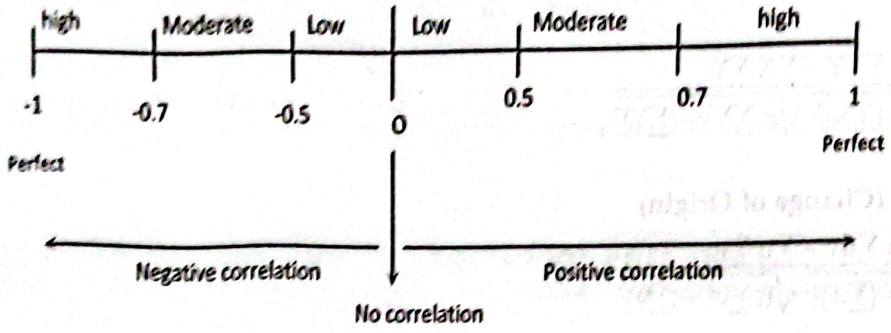
v. If r lies between 0.5 to 0.699 there is moderate positive correlation

vi. If r lies between 0.70 to 0.999 High degree positive correlation

vii. If r lies between -0.499 to -0.001, there is Low degree of Negative Correlation

viii. If r lies between -0.699 to -0.500 there is moderate Negative Correlation.

ix. If r lies between -0.699 to -0.999 there is high degree of Negative Correlation.



Example 1: **If the covariance between X and Y is 27 and the variance of X is 25 and the variance of Y is 64, find the Karl Pearson's correlation coefficient.**

Solution

Here, Cov (x,y) = 27, Var(c) = 25 and Var(y) = 64

Karl Pearson's correlation coefficient (r) $= \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}\sqrt{\text{Var}(y)}} = \frac{27}{\sqrt{25}\sqrt{64}} = \frac{27}{5 \times 8} = 0.675$

Example 2: **The coefficient of correlation between two variable x and y is 0.38. Their covariance is 10.2. The variance of x is 16, find the standard deviation of y series.**

Solution

Here, r = 0.38, Cov(x,y) = 10.2, Var(x) = 16, SD(y) = ?

We know

$$r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)}\ \sqrt{\text{Var}(y)}}$$

or $0.38 = \frac{10.2}{\sqrt{16}\sqrt{\text{Var}(y)}}$

or $0.38 = \frac{10.2}{4\sqrt{\text{Var}(y)}}$

or $\sqrt{\text{Var}(y)} = \frac{10.2}{4 \times 0.38} = 6.71$

or $\text{SD}(y) = 6.71$

Example 3: **Calculate the Correlation Coefficient between x and y series from the following data:**

	x series	y series
No of pair of observation	5	5
Arithmetic mean	6	8
Standard deviation	2.82	2
Sum of square of deviations from mean	40	20

Sum of the product deviation of x and y series from their respect mean = -26

Solution:

Here, $n = 5$, $\bar{X} = 6$, $\bar{Y} = 8$, $\sigma_X = 2.82$, $\sigma_Y = 2$, $\Sigma(X - \bar{X})^2 = 40$, $\Sigma(Y - \bar{Y}) = 20$, $\Sigma(X - \bar{X})(Y - \bar{Y}) = -26$

$$\therefore \quad r = \frac{\Sigma(X - \bar{X})(y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2}\sqrt{\Sigma(Y - \bar{Y})^2}} = \frac{-26}{\sqrt{40}\sqrt{20}} = -0.9192$$

Example 4: **For 10 pair of observations on two variables X and Y,the following information are obtained; $\Sigma x = 666$, $\Sigma y = 663$, $\Sigma x^2 = 44490$, $\Sigma y^2 = 44061$, $\Sigma xy = 44224$. Compute Karl Pearson's correlation coefficient.**

Solution

$$\text{Karl Pearson's correlation coefficient (r)} = \frac{n\Sigma XY - n\Sigma x \Sigma Y}{\sqrt{n(\Sigma X^2 - (\Sigma X)^2)}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10 \times 44224 - 666 \times 663}{\sqrt{10 \times 44490 - (666)^2}\sqrt{10 \times 44061 - (663)^2}}$$

$$= \frac{682}{36.66 \times 32.264} = 0.576$$

Example 5: **Calculate the Karl Pearson's correlation coefficient between number of pages in a book and number of mistakes.**

Number of pages of book X	6	2	10	4	8
Number of mistakes Y	9	11	5	8	7

Solution:

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

Here,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6, \bar{Y} = \frac{\Sigma Y}{n} = \frac{40}{5} = 8, x = X - \bar{X}, y = Y - \bar{Y}$$

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{-26}{\sqrt{40}\sqrt{20}} = -0.919 > 0.$$

High degree negative correlation

Example 6: **Calculate Karl Pearson's correlation coefficient between sales and repair computers as given below.**

Sales	50	55	55	60	65	70	65	60
Repair	11	13	14	16	16	15	15	20

Solution

Sales(X)	Repair(Y)	u = X - A	v = Y - B	uv	u^2	v^2
50	11	-15	-5	75	225	25
55	13	-10	-3	30	100	9
55	14	-10	-2	20	100	4
60	16	-5	0	0	25	0
65	16	0	0	0	0	0
70	15	5	-1	-5	25	1
65	15	0	-1	0	0	1
60	20	-5	4	-20	25	16
$\Sigma X = 45$	$\Sigma Y = 108$	$\Sigma u = 597$	$\Sigma v = 285$	$\Sigma uv = 1356$	$\Sigma u^2 = 500$	$\Sigma v^2 = 56$

Let A = 65, B = 16

$$\text{Correlation coefficient (r)} = \frac{n\Sigma uv - \Sigma u.\,\Sigma v}{\sqrt{n\,\Sigma u^2 - (\Sigma u)^2}\sqrt{n\,\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{8 \times 100 - (-40) \times (-8)}{\sqrt{8 \times 500 - (-40)^2}\sqrt{8 \times 56 - (-8)^2}} = \frac{800 - 320}{48.990 \times 16.596}$$

$$= 0.50$$

Example 7: **The following table gives the distribution of items and also defective items among them according to size groups. Find the correlation coefficient between size and defect in quality.**

Size Group	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45
No. of items	200	270	340	360	400	300
No of defective items	150	162	170	180	180	120

Solution:

Since correlation is to be found between size and defect in quality.

No. of items	No. of defect in items	Percentage of defect item
200	150	$\frac{150}{200} \times 100 = 75$
270	162	$\frac{162}{270} \times 100 = 60$
340	170	$\frac{170}{340} \times 100 = 50$
360	180	$\frac{180}{360} \times 100 = 50$
400	180	$\frac{180}{400} \times 100 = 45$
300	120	$\frac{120}{300} \times 100 = 40$

Let X and Y be the mid value of age and percentage of number of players respectively.

Age group	Mid value (X)	(Y)	$u' = \frac{X - 27.5}{5}$	$v' = \frac{Y - 50}{5}$	$u'v'$	u'^2	v'^2
15 - 20	17.5	75	-2	5	-10	4	25
20 - 25	22.5	60	-1	2	-2	1	4
25 - 30	27.50	50	0	0	0	0	0
30 - 35	32.5	50	1	0	0	1	0
35 - 40	37.5	45	2	-1	-2	4	1
40 - 45	42.5	40	3	-2	-6	9	4
			$\Sigma u'=3$	$\Sigma v'=4$	$\Sigma u'v'=-20$	$\Sigma u^2=19$	$\Sigma v'^2=34$

Now,

$$r = \frac{n\Sigma u'v' - \Sigma u'.\Sigma v'}{\sqrt{n\Sigma u'^2 - (\Sigma u')^2}\sqrt{n\Sigma v'^2 - (\Sigma v')^2}}$$

$$= \frac{6 \times (-20) - 3 \times 4}{\sqrt{6 \times 19 - (3)^2}\sqrt{6 \times 34 - (4)^2}}$$

$$= \frac{-120 - 12}{\sqrt{105 \times 188}} = -0.94$$

Example 8: **In order to find the correlation coefficient two variables x and y from 12 observations, the following calculations were made $\Sigma X = 30$, $\Sigma Y = 5$, $\Sigma X^2 = 670$, $\Sigma Y^2 = 285$, $\Sigma XY = 334$.**

On checking data it was found that the pair (X = 11, Y = 4) was copied wrongly, the correct value being (X = 10, Y = 14). Find the correct value of correlation coefficient.

Solution:

Here $n = 12$, $\Sigma X = 30$, $\Sigma X^2 = 670$, $\Sigma Y^2 = 285$, $\Sigma XY = 334$

Now Corrected $\Sigma X = 30 - 11 + 10 = 29$

Corrected $\Sigma Y = 5 - 4 + 14 = 15$

Corrected $\Sigma X^2 = 670 - 11^2 + 10^2 = 649$

Corrected $\Sigma Y^2 = 285 - 4^2 + 14^2 = 465$

Corrected $\Sigma XY = 334 - 11 \times 4 + 10 \times 14 = 430$

Then

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

$$= \frac{12 \times 430 - 29 \times 15}{\sqrt{12 \times 649 - (29)^2}\sqrt{12 \times 465 - (15)^2}} = \frac{51 - 435}{\sqrt{7788 - 841}\sqrt{5580 - 225}} = \frac{4725}{\sqrt{6947}\sqrt{5355}} = 0.7757$$

High Degree of Positive Correlation.

Goodness of Fit measure in terms of Correlation Coefficient

Probable error is a statistical measure for testing the reliability of the value of Correlation Coefficient. It is used to test the calculated correlation coefficient whether it is significant or not. If r is calculated correlation coefficient in a sample of n pair of observation then its standard error as $S.E.(r) = \frac{1-r^2}{\sqrt{n}}$ and Probable error is defined as

$$P.E.(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

Reason of taking the factor 0.6745 is that in a normal distribution 50% of the observations lie in this range.

- If $r < P.E.(r)$, then the value of r is not significant.
- If $r > 6\ P.E.(r)$, then the value of r is significant.

In other situation, nothing can be concluded certainly. The probable error of correlation is used to determine the limits within which the population correlation coefficient may be expected to lie i.e. limits of population correlation coefficient = $r \pm P.E.(r)$.

Example 9: **If the correlation coefficient between X and Y is 0.7, the probable error of r is 0.0344, what will be the value of n?**

Solution:

Here, $r = 0.7$, $P.E.(r) = 0.0344$, $n = ?$

We know

$$P.E.(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

or $$0.0344 = 0.6745 \times \frac{1 - (0.7)^2}{\sqrt{n}}$$

or $$\sqrt{n} = \frac{0.6745 \times 0.51}{0.0344}$$

or $$\sqrt{n} = 9.999$$

or $$n = 99.98 \approx 100$$

Example 10: **The coefficient of correlation between 16 pairs of values of work hours and salary was found to be 0.7. Test the significance of the correlation coefficient.**

Solution:

Here, $n = 16$, $r = 0.7$

To test the significance of correlation coefficient,

$$P.E.(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

$$= 0.6745 \times \frac{1 - 0.7^2}{\sqrt{16}} = 0.6745 \times \frac{0.51}{4} = 0.085$$

Here $r = 0.7$ is not less than $P.E.(r) = 0.085$

$6\ P.E.(r) = 6 \times 0.085 = 0.515$

Now,

$r = 0.7 > 6\ P.E.(r) = 0.515$

The correlation coefficient between work hours and salary is found to be significant.

Example 11: **Find the correlation between x and y for given frequency as follows. And test consistency of x and y.**

x	y	f
10	24	12
12	28	24
14	19	18
16	15	10
18	12	18
20	13	5

Solution:

x	y	f	fx	fx^2	fy	fy^2	fxy
10	24	12	120	1200	288	6912	2880
12	28	24	288	3456	672	18816	8064
14	19	18	252	3528	342	6498	4788
16	15	10	160	2560	150	2250	2400
18	12	18	324	5832	216	2592	3888
20	13	5	100	2000	65	845	1300
		Σf=87	Σfx = 1244	Σfx^2= 18576	Σfy = 1733	Σfy^2 = 37913	Σfxy = 23320

Here, Total frequency = N = 87

Mean of $x = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{1244}{87} = 14.299$

Mean of $y = \bar{y} = \frac{\Sigma fy}{\Sigma f} = \frac{1733}{87} = 19.919$

Standard deviation of $x = \sigma_x = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \left(\frac{\Sigma fx}{\Sigma f}\right)^2} = 3.01$

Standard deviation of $y = \sigma_y = \sqrt{\frac{\Sigma fy^2}{\Sigma f} - \left(\frac{\Sigma fy}{\Sigma f}\right)^2} = 6.244$

Coefficient of variation of x = C.V. (x) $= \frac{\sigma_x}{\bar{x}} \times 100\% = 21.05\%$

Coefficient of variation of y = C.V. (y) $= \frac{\sigma_y}{\bar{y}} \times 100\% = 31.348\%$

Karl Pearsons' correlation coefficient

$$r = \frac{N\Sigma fxy - \Sigma fx\,\Sigma fy}{\sqrt{N\Sigma fx^2 - (\Sigma fx)^2}\sqrt{N\Sigma fy^2 - (\Sigma fy)^2}} = -0.892$$

Example 12: **A student calculates the value of correlation coefficient between study hour and marks secured is 0.795 when the number of items is 100 and concludes that r is highly significant. Is the conclusion correct? Also determine the limit of population correlation coefficient.**

Solution:

Here, $r = 0.795$, $n = 100$

To test the significance of correlation coefficient,

$$P.E.(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

$$= 0.6745 \times \frac{1 - 0.795^2}{\sqrt{100}}$$

$$= 0.6745 \times \frac{0.3679}{10}$$

$$= 0.0248$$

Here $r = 0.795$ is not less than $P.E.(r) = 0.0248$

$6\ P.E.(r) = 6 \times 0.0248 = 0.1489$

$r = 0.795 > 6\ P.E.(r) = 0.1489$

Hence the conclusion that r is highly significant is correct.

Now, limit of population correlation coefficient $= r \pm P.E.(r)$

$= 0.795 \pm 0.0248$

Taking - sign

$0.795 - 0.0248 = 0.7702$

Taking + sign

$0.795 + 0.0248 = 0.8198$

Hence, the limit of the population correlation coefficient is 0.7702 to 0.8198.

Example 13: **Find whether there exists any relationship between the age of driver and no of accidents from the following table.**

No of accidents	Age of driver				
	20 - 22	22 - 24	24 - 26	26 - 28	28 - 30
0	5	7	13	-	-
1	-	10	15	2	-
2	-	3	12	6	-
3	-	-	4	12	5
4	-	-	-	4	2

Solution:

$$u = x - 2,\ v = \frac{y - 25}{2}$$

Let X and Y represent age of driver and no of accidents

	Age of driver	20 - 22	22 - 24	24 - 26	26 - 28	28 - 30				
	Mid value (Y)	21	23	25	27	29				
No of accidents (X)	v / u	-2	-1	0	1	2	f	fu	fu^2	fuv
0	-2	5 [20]	7 [14]	13 [0]	-	-	25	-50	100	34
1	-1	-	10 [10]	15 [0]	2 [-2]	-	27	-27	27	8
2	0	-	3 [0]	12 [0]	6 [0]		21	0	0	0
3	1	-		4 [0]	12 [12]	5 [10]	21	21	21	22
4	2	-			4 [8]	2 [8]	6	12	24	16
	f	5	20	44	24	7	N = 100	Σfu = -44	$Σfu^2$ = 172	Σfuv = 80
	fv	-10	-20	0	24	14	Σfv =8			
	fv^2	20	20	0	24	28	$Σfv^2$ =92			
	fuv	20	24	0	18	18	Σfuv = 80			

$$r = \frac{n\Sigma fuv - \Sigma fu.\,\Sigma fv}{\sqrt{n\,\Sigma fu^2 - (\Sigma fu)^2}\sqrt{n\,\Sigma fv^2 - (\Sigma fv)^2}}$$

$$= \frac{100 \times 80 - 8 \times (-44)}{\sqrt{100 \times 92 - (8)^2}\sqrt{100 \times 172 - (-44)^2}} = \frac{800 + 352}{\sqrt{9200 - 64}\sqrt{17200 - 1936}}$$

$$= \frac{8352}{95.58 \times 123.55} = 0.71$$

There exists high degree of positive correlation between age of driver and number of accident.

REGRESSION ANALYSIS

The term regression means stepping back towards the average. It is a statistical tool used to determine the nature of relationship that exists among two or more variables and making estimate or prediction from that relationship. The unknown variable that we are going to estimate is called dependent variable or explained variable or regressed and the known variable is called independent variable or regressor or explanatory variable. For example, let there is a high correlation between day temperature and sales of cold drinks then the salesman of cold drinks might wish to know the forecasted temperature for the next day to decide for the stock of cold drinks. This can be done with the help of regression. When two variable x and y are highly correlated one can find the best estimated value of x for given value of y or best estimated value of y for given value of x.

The principle difference of correlation analysis and regression analysis is that the correlation analysis indicates to what extent the variables are related while the regression analysis indicates how the variables are related.

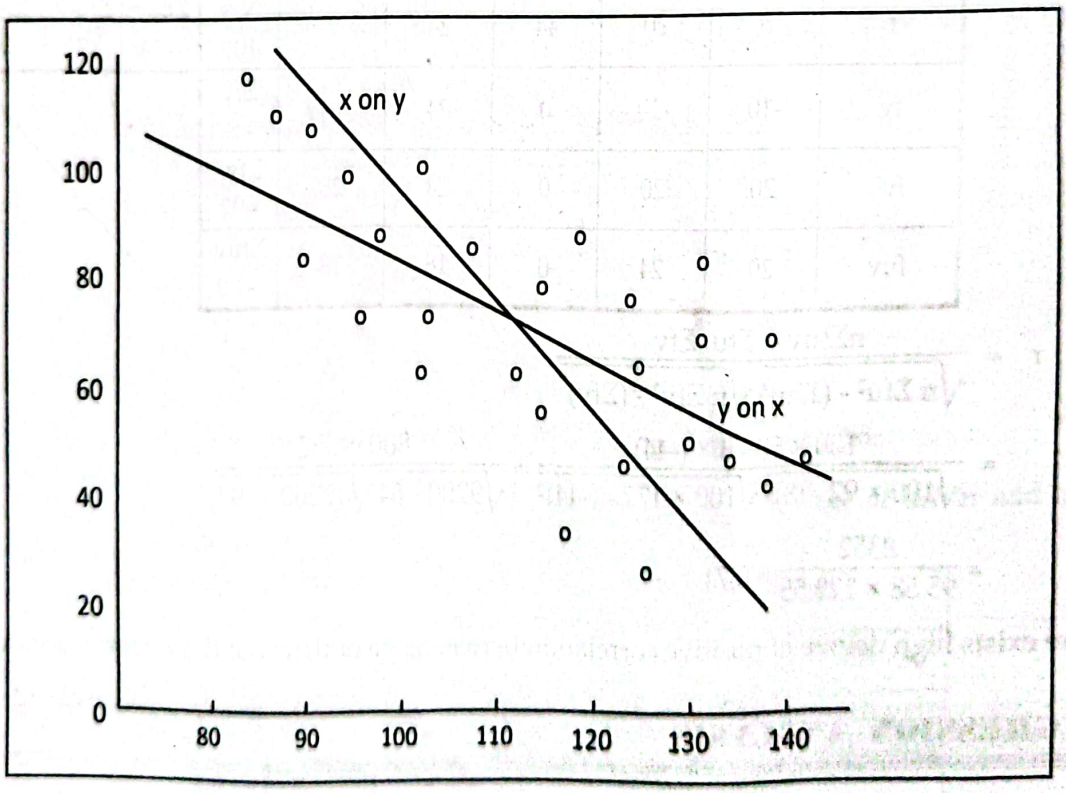
Linear and Non-linear Regression

When the graph between dependent and independent variable is linear trend then the regression is called linear regression. When the graph between dependent and independent variable is not linear trend then the regression is called nonlinear regression.

Lines of Regression

In case of two variables x and y we have two regression lines. One is regression line of y on x and other is regression line of x on y. The regression line of y on x estimates the value of y for given value of x. The regression line of x on y estimates the value of x for given value of y.

If there is perfect correlation($r = \pm 1$) between two variables the two regression lines will coincide with each other. The regression lines are farther from each other as lesser degree of correlation between the variables. The regression lines are nearer to each other as high degree of correlation between the variables. In case of independent variables regression lines will be perpendicular to each other.



Regression Equations

The algebraic expressions of the regression lines are called regression equations. For two variables in which one is independent and other is dependent variable there are two regression lines. Hence there are two regression equations.

i. Regression equation of y on x given by $y = a + bx$ in which y is dependent variable and x is independent variable.

ii. Regression equation of x on y given by $x = a + by$ in which x is dependent variable and y is independent variable.

Regression Equation of y on x

Let us suppose in bivariate distribution (x_i, y_i), i = 1,2,3...n dependent variable is y and independent variable is x. Regression equation of y on x is $y = a+bx$

Here a and b are constants and are estimated by using the principle of least square by minimizing error sum of square.

Simple Linear Regression

It is the linear function of a dependent variable with independent variable. With the help of independent variable the values of dependent variable can be predicted. It is statistical tool for finding the linear relationship of a dependent variable with independent variable. If the variables in bivariate distribution are related in linear form then the regression is linear otherwise curvilinear. It gives the best estimate to the value of one variable for any specific value of other variable.

Let us consider bivariate distribution (x_i, y_i); i = 1,2,3,.....n; y is dependent variable and x is independent variable then regression equation of y on x is given by $y = a+bx$, where a and b are constants and called y intercept and regression coefficient or slope respectively. b measures amount of change in y per unit change in x.

Assumptions of Linear Regression

Let us consider multiple regression model

$y = a + bx + \varepsilon$

There are certain assumptions about the model. The assumptions are based on relation between error ε and explanatory variables x

The following are the major assumptions on the random errors ε which is considers a serious problem in modeling if any of them is violated by the error term.

i. Regression model is linear in parameters.
ii. ε is random real variable
iii. The random errors ε have zero mean, i.e.,$E(\varepsilon) = 0$
iv. The random errors ε has constant variance i.e., $V(\varepsilon) = \sigma_e^2$ (No heteroscedaticity).
v. The random variable ε is normally distributed. i.e., $\varepsilon \sim N(0, \sigma^2)$
vi. The random errors ε are independent i.e., $E(\varepsilon_i\varepsilon_j) = 0 : i \neq j$. (No autocorrelation).
vii. xs are uncorrelated to the error term ε, i.e., $E(X\varepsilon) = 0$ (uniformity of x over samples
viii. The explanatory variable x measured without error.

ESTIMATION OF COEFFICIENTS USING LEAST SQUARE METHOD (OLS)

The values of a and b are determined by using the principle of least square by minimizing error sum of square.

Here, error (e) = $y - \hat{y}$ so that $\Sigma e^2 = \Sigma(y - \hat{y})^2$

Let $S = \Sigma e^2 = \Sigma(y-\hat{y})^2 = \Sigma(y - a - bx)^2$

Differentiating both sides with respect to a

$$\frac{d\Sigma e^2}{da} = \frac{d\Sigma(y - a - bx)^2}{da} = 2\Sigma(Y - a - bx)(-1) = -2\Sigma(y - a - bx)$$

For S to be minimum

$$\frac{d\Sigma e^2}{da} = 0$$

or $-2\Sigma(y - a - bx) = 0$

or $\Sigma(y - a - bx) = 0$

or $\Sigma y - na - b\Sigma x = 0$

or $\Sigma y = na + b\Sigma x$ .....(i)

Differentiating both sides with respect to b

$$\frac{d\Sigma e^2}{db} = \frac{d\Sigma(y - a - bx)^2}{db} = 2\Sigma(y - a - bx)(-x) = -2\Sigma(yx - ax - bx^2)$$

For S to be minimum

$$\frac{d\Sigma e^2}{db} = 0$$

or $-2\Sigma(yx - ax - bx^2) = 0$

or $\Sigma(yx - ax - bx^2) = 0$

or $\Sigma yx - a\Sigma x - b\Sigma x^2 = 0$

or $\Sigma yx = a\Sigma x + b\Sigma x^2$ .....(ii)

Solving normal equations i and ii get a and b and substitute value to get the regression equation $y = a + bx$

In equation $Y = a + bx$; a = intercept,

b = regression coefficient (change in y per unit change in x) = slope

Linear Relationship

A relationship of direct proportionality that, when plotted on a graph, traces a straight line. In linear relationships, any given change in an independent variable will always produce a corresponding change in the dependent variable.

For example, a linear relationship between production hours and output in a factory means that a 10 percent increase or decrease in hours will result in a 10 percent increase or decrease in the output.

Deviation method

$u = x - A$ and $v = y - B$

Obtain regression equation of v on u i.e.,$v = a+bu$

Obtain a and b using principle of least square by minimizing error sum of square solving

$\Sigma v = na + b\Sigma u$ ..... (i)

$\Sigma uv = a\Sigma u + b\Sigma u^2$ ..... (ii)

Substitute value of a and b in $v = a+bu$ to get equation of v on u

Substitute value of $u = x - A$ and $v = y - B$ to get equation $y = a+bx$

Step Deviation Method

$$u' = \frac{x - A}{h} \text{ and } v' = \frac{y - B}{k}$$

Obtain regression equation of v' on u' i.e., $v = a+bu'$

Obtain a and b using principle of least square by minimizing error sum of square solving

$$\Sigma v' = na + b\Sigma u' \quad \ldots\ldots (i)$$

$$\Sigma u'v' = a\Sigma u' + b\Sigma u'^2 \quad \ldots\ldots (ii)$$

Substitute value of a and b in $v' = a+bu$ to get equation of v' on u'

Substitute value of $u' = \frac{x - A}{h}$ and $V' = \frac{y - B}{k}$ to get equation $y = a + bx$

For regression equation of x on y

Equation be $x = a + by$

To estimate a and b using principle of least square by minimizing error sum of square

$$\Sigma x = na + b\Sigma y \quad \ldots\ldots (i)$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2 \quad \ldots\ldots (ii)$$

Solving (i) and (ii) get a and b and substitute in $x = a+by$

Deviation method

$u = x - A$ and $v = y - B$

Obtain regression equation of u on v i.e.,$u = a+bv$

Obtain a and b using principle of least square by minimizing error sum of square solving

$$\Sigma u = na + b\Sigma v \quad \ldots\ldots (i)$$

$$\Sigma uv = a\Sigma v + b\Sigma v^2 \quad \ldots\ldots (ii)$$

Substitute value of a and b in $u = a+bv$ to get equation of u on v

Substitute value of $u = x - A$ and $v = y - B$ to get equation $x = a+by$

Step deviation method

$$u' = \frac{x - A}{h} \text{ and } v' = \frac{y - B}{k}$$

Obtain regression equation of u' on v'i.eu' = a+bv'

Obtain a and b using principle of least square by minimizing error sum of square solving

$$\Sigma u' = na + b\Sigma v' \quad \ldots\ldots (i)$$

$$\Sigma u'v' = a\Sigma v' + b\Sigma v'^2 \quad \ldots\ldots (ii)$$

Substitute value of a and b in $u' = a+bv'$ to get equation of u' on v'

Substitute value of $u' = \frac{x - A}{h}$ and $v' = \frac{y - B}{k}$ to get equation $x = a+by$

Properties of Regression Coefficients

i) Correlation coefficient is the geometric mean between the regression coefficients.
$r = (b_{YX} \times b_{XY})^{1/2}$

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

i) Product of two regression coefficients must be less than or equal to 1.

ii) Arithmetic mean of regression coefficients is greater than the correlation coefficient.

iii) Regression coefficients are independent of change of origin but not of scale.

iv) In regression equation of y on x, $y = a+bx$, b is slope of the line and is called regression coefficient of y on x. It measures change in dependent variable y per unit change in independent variable x. It is given by

$$b_{yx} = \frac{\text{Covriance }(x, y)}{\text{Variance }(x)} = \frac{r\sigma_x\sigma_y}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

Example 14: **For the 10 observations, following information were obtained: $\Sigma X = 130$, $\Sigma Y = 220$, $\Sigma X^2 = 2$ [illegible], $\Sigma Y^2 = 5506$, $\Sigma XY = 3467$.**

Obtain the line of regression of Y on X and estimate Y when X is 16?

Solution:

Here, Y= ? When X = 16

We know

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

or, $220 = 10a+130b$ .... (i)

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

or, $3467 = 130a+ 2238b$ .... (ii)

Multiply (i) by 13 and subtract from (ii)

$3467 - 2860 = 130a+2238b - 130a - 1690b$

or, $607 = 548b$

or, $b = 1.109$

Substitute b in (i)

$220 = 10a + 130\times 1.109$

or, $220 = 10a+144.17$

or, $10a = 220 - 144.17$

or, $a = \frac{75.83}{10} = 7.583$

Regression equation of Y on X is $Y = a+bX$

or, $Y = 7.583+ 1.109X$

When X= 16, $Y = 7.583 + 1.109 \times 16 = 25.32$

Example 15: **Author believes that there is a linear relationship between verbal test score of students(Y) and the number of library books checked out(X). Following are the data collected on 10 students.**

X	12	15	3	7	10	5	22	9	13	7
Y	77	85	48	59	75	41	94	65	79	70

The above data reveals the following results;

$\Sigma X = 103$, $\Sigma Y = 693$, $\Sigma X^2 = 1335$, $\Sigma Y^2 = 50447$, $\Sigma XY = 7881$. Fit a simple linear regression model of Y on X. Interpret the slope regression coefficient

Solution:

To fit $Y = a+bX$

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

or $693 = 10a+103b$ ....(i)

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

or $7881 = 103a + 1335b$ ....(ii)

Coeff. of a	Coeff. of b	Constant
10	103	693
103	1335	7881

$$D = \begin{vmatrix} 10 & 103 \\ 103 & 1335 \end{vmatrix} = 13350 - 10609 = 2741$$

$$D_1 = \begin{vmatrix} 693 & 103 \\ 7881 & 1335 \end{vmatrix} = 925155 - 811743 = 113412$$

$$D_2 = \begin{vmatrix} 10 & 693 \\ 103 & 7881 \end{vmatrix} = 78810 - 71379 = 7431$$

Now, $a = \frac{D_1}{D} = \frac{113412}{2741} = 41.37$

$b = \frac{D_2}{D} = \frac{7431}{2741} = 2.711$

Regression equation of y on x

$y = a+bx = 41.37 + 2.711x$

Here b = 2.711 it means y changes by 2.711 per unit change in x.

Example 16: **From the following data obtain the regression equation current in amperes and resistance in ohms using the method of least square and estimate the probable value of current if resistance is 4 ohms.**

Resistance in ohms	2	3	5	6	8
Current in amperes	10	8	7	3	1

Solution:

Resistance in ohms x	Current in amperes y	x^2	y^2	xy
2	10	4	100	20
3	8	9	64	24
5	7	25	49	35
6	3	36	9	18
8	1	64	1	8
$\Sigma x = 24$	$\Sigma y = 29$	$\Sigma x^2 = 138$	$\Sigma y^2 = 223$	$\Sigma xy = 105$

To fit $y = a + bx$

$\Sigma y = na + b\Sigma x$

or $29 = 5a + 24b$ .....(i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

or $105 = 24a + 138b$ .....(ii)

Multiply equation i by 24 and ii by 5 and subtract

$696 = 120a + 576b$

$\underline{-\ 525 = -\ 120a + -\ 690b}$

$171 = 0 - 114b$

or, $b = \frac{-171}{114} = -1.5$

Substitute b in i

$29 = 5a + 24(-1.5)$

or, $29 + 36 = 5a$

or, $a = \frac{65}{5} = 13$

Regression equation of y on x is $y = a + bx$

or $y = 13 + (-1.5)x = 13 - 1.5x$

When $x = 4$, $y = 13 - 1.5 \times 4 = 7$

Hence value of y is 7 when value of x is 4.

Example 17: **The following table gives normal weight of a baby during the first six months of life.**

Age in months	0	2	3	5	6
Weight in lbs	5	7	8	10	12

Determine the regression equation of weight on age using method of least squares and estimate the weight of body at the age of 4 months.

Solution

Age in months (x)	Weight in lbs (y)	xy	x^2
0	5	0	0
2	7	14	4
3	8	24	9
5	10	50	25
6	12	72	36
$\Sigma x = 16$	$\Sigma y = 42$	$\Sigma xy = 160$	$\Sigma x^2 = 74$

To fit $y = a + bx$

$\Sigma Y = na + b\Sigma x$

or $42 = 5a + 16b$ .....(i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

or $160 = 16a = 74b$ .....(ii)

From equation i and ii

Coeff of a	Coeff of b	Constant
5	16	42
16	74	160

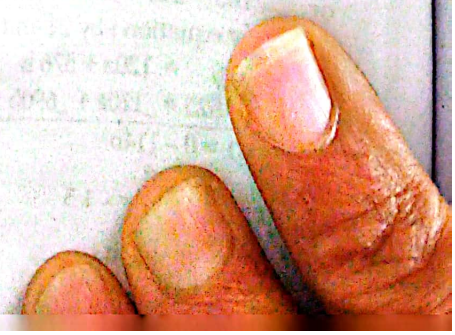
Now $D = \begin{vmatrix} 5 & 16 \\ 16 & 74 \end{vmatrix} = 114$

$D_1 = \begin{vmatrix} 42 & 16 \\ 160 & 74 \end{vmatrix} = 548$

$D_2 = \begin{vmatrix} 5 & 42 \\ 16 & 160 \end{vmatrix} = 128$

Now,

$a = \frac{D_1}{D} = \frac{548}{114} = 4.8 \quad b = \frac{D_2}{D} = \frac{128}{114} = 1.122$



Hence the regression equation is $y = a + bx = 4.8 + 1.122x$

When age $(x) = 4$

$y = 4.8 + 1.122 \times 4 = 9.288$

Hence estimated body weight is 9.288 lbs at age of 4 months.

Example 18: **From the following data between yearly turnover and profits. Find regression equation of profit and yearly turnover and estimate profit when yearly turnover is 30 million units.**

Profit in thousand $	18	20	22	23	27	28	30
Yearly turnover in million units	23	25	27	30	32	31	35

Solution:

Profit in thousand $ (y)	Yearly turnover in million units(x)	v = y - 23	u = x - 30	uv	u^2
18	23	-5	-7	35	49
20	25	-3	-5	15	25
22	27	-1	-3	3	9
23	30	0	0	0	0
27	32	4	2	8	4
28	31	5	1	5	1
30	35	7	5	35	25
		$\sum v = 7$	$\sum u = -7$	$\sum uv = 101$	$\sum u^2 = 113$

First fit equation of v on u

$v = a + bu$

$\sum v = na + b\sum u$

or $7 = 7a - 7b$ .....(i)

$\sum uv = a\sum u + b\sum u^2$

or $101 = -7a + 113b$ .....(ii)

Add (i) and (ii)

$7 + 101 = 7a - 7b - 7a + 113b$

or $108 = 106b$

or $b = \frac{108}{106} = 1.018$

Substitute b in (i)

$7 = 7a - 7 \times 1.018$

or $7 = 7a - 7.13$

or $a = \frac{14.13}{7} = 2.01$

Now, $v = a + bu$

or $y - 23 = 2.10 + 1.018(x - 30)$

or $y = 2.1 + 1.018x - 33.24 + 23$

or $y = -8.14 + 1.018x$

Regression equation of profit with yearly turnover is $y = -8.14 + 1.018x$

When yearly turnover(x) = 30,

$y = -8.14 + 1.018 \times 30 = 22.31 \approx 22$

Hence estimated profit is \$22,000 when yearly turnover is 30 million units.

Example 19: Find the regression equation of X on Y from following data.

X	5	15	20	25	30
Y	50	60	80	110	130

Solution:

X	Y	$u' = \frac{x-20}{5}$	$v' = \frac{y-80}{10}$	$u'v'$	v'^2
5	50	-3	-3	9	9
15	60	-1	-2	2	4
20	80	0	0	0	0
25	110	1	3	3	9
30	130	2	5	10	25
		$\Sigma u' = -1$	$\Sigma v' = 3$	$\Sigma u'v' = 24$	$\Sigma v'^2 = 47$

Now,

Regression equation of u' on v' is

$u' = a + bv'$

$\sum u' = na + b\sum v'$

or $-1 = 5a + 3b$ .....(i)

$\sum u'v' = a\sum v' + b\sum v'^2$

or $24 = 3a + 47b$ .....(ii)

Coeff. of a	Coeff. of b	Constant
5	3	-1
3	47	24

$D = \begin{vmatrix} 5 & 3 \\ 3 & 47 \end{vmatrix} = 235 - 9 = 226$

$D_1 = \begin{vmatrix} -1 & 3 \\ 24 & 47 \end{vmatrix} = -47 - 72 = -119$

$D_2 = \begin{vmatrix} 5 & -1 \\ 3 & 24 \end{vmatrix} = 120 + 3 = 123$

Now, $a = \frac{D_1}{D} = -\frac{119}{226} = -0.526$

$b = \frac{D_2}{D} = \frac{123}{226} = 0.544$

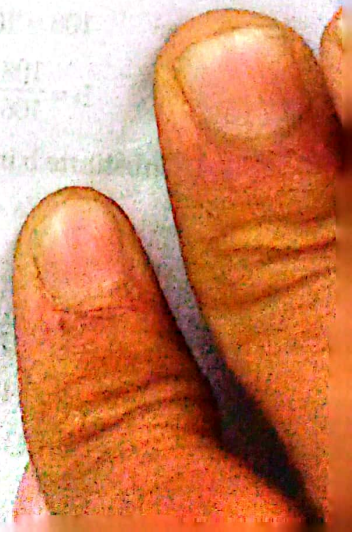
Regression equation of u' on v'

$u' = a + bv'$

$u' = -0.526 + 0.554\, v'$

or $\frac{x-20}{5} = -0.526 + 0.554\left[\frac{y-80}{10}\right]$

or $\frac{x-20}{5} = \frac{-5.26 + 0.554\, y - 44.32}{10}$



or $x - 20 = \frac{-49.58 + 0.554y}{2}$

or $2x - 40 = -49.58 + 0.554y$

or $2x = -9.58 + 0.554y$

or $x = -4.79 + 0.277y$

The required equation of x on y is $x = -4.79 + 0.272y$

Difference between correlation and regression

Correlation	Regressior
1. It measures the degree to which the variables are linearly related.	1. It measures the average relationship between variables whether variables are linearly related or not.
2. It is symmetric i.e. $r_{yx} = r_{xy}$	2. It is not symmetric. i.e. $b_{YX} \neq b_{XY}$.
3. It need not imply cause and effect relationship between the variables under study.	3. It clearly indicate the cause and effec relationship between the variables under study.
4. It is pure number independent of unit of measurement.	4. It is not pure number attached with the unit of measurement.
5. It has limited application as compared to regression.	5. It has wide applications.
6. It is relative measure.	6. It is absolute measure.
7. There may be non-sense correlation.	7. There may not be non-sense regression.

MEASURES OF VARIATION

In regression model value of dependent variable are estimated on the basis of independent variables. In regression analysis total variation is divided into explained variation (sum of square due to regression) and unexplained variation (sum of square due to error). Hence according to Fisher total sum of square is decomposed into sum of square due to regression and sum of square due to error.

Total sum of square (TSS) = Sum of square due to regression (SSR) + Sum of square due to error (SSE)

For the regression model Y= a + bX, where Y is dependent variable and X is independent variable.

$TSSS = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$ = Total variation

$SSE = (y - \hat{y})^2$

It is also given by

$SSE = \Sigma y^2 - a\Sigma y - b\Sigma xy$ = Unexplained variation

SSR = TSS – SSE = Explain variation.

STANDARD ERROR OF THE ESTIMATE

Standard error is the Square root of the variance computed from sample data. The standard error of the estimate measures the average variation or scatterness of the observed data point around regression line. Standard error of the estimate is used to measure the reliability of the regression equation. Regression line having less standard error of estimate is more reliable than regression line having more standard error of estimate.

It is given by $S_e = \sqrt{\frac{SSE}{n-k-1}}$

SSE = sum of square due to error

k = number of independent variables in regression model

n = number of observations.

When $S_e = 0$, there is no variation of observed data around regression line. In such case regression line is perfect for estimating the dependent variable.

COEFFICIENT OF DETERMINATION

It measures the proportion of variation in dependent variable that is explained by the set of independent variables. It is the measure based upon measure of variation and is used to determine the fitness of the data to the model. The regression line is reliable if the sum of square due to regression is much greater than sum of square due to error. It is the ratio of sum of square due to regression to the total sum of square. It is denoted by R^2 and is given by, $R^2 = \frac{SSR}{TSS}$.

For regression equation of y on x

$$TSS = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2$$

$$SSE = \Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - a\,\Sigma Y - b\,\Sigma XY$$

$$SSR = TSS - SSE$$

It is also obtained by simply squaring the correlation coefficient. i.e. $R^2 = r^2$. Higher the value of R^2 the more reliable is the fitted equation. It lies between 0 and 1.

R^2 can never decrease when another independent variable is added to a regression. R^2 will usually increase with increase in number of independent variables.

Example 20: **Calculate the coefficient of determination for following data.**

X	50	55	55	60	65	70	65	60
Y	11	13	14	16	16	15	15	20

Solution

X	Y	X^2	Y^2	XY
50	11	2500	121	550
55	13	3025	169	715
55	14	3025	196	770
60	16	3600	256	960
65	16	3600	256	1040
70	15	4900	225	1050
65	15	4225	225	975
60	20	3600	400	1200
$\Sigma X = 480$	$\Sigma Y = 120$	$\Sigma X^2 = 29100$	$\Sigma Y^2 = 1848$	$\Sigma XY = 7260$

We have,

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

$$= \frac{8\times7260 - 480\times120}{\sqrt{8\times29100 - (480)^2}\sqrt{8\times1848 - (120)^2}}$$

$$= \frac{480}{\sqrt{232800 - 230400}\sqrt{14787 - 14400}}$$

$$= \frac{480}{48.989\times19.596} = 0.50$$

Now coefficient of determination $R^2 = r^2 = (0.5)^2 = 0.25$

Example 21: From 5 pair of observations it was found that $\Sigma X = 16$, $\Sigma Y = 42$, $\Sigma XY = 160$, $\Sigma Y^2 = 382$ and $\Sigma X^2 = 74$. Find coefficient of determination.

Solution:

We know

$$TSS = \Sigma(Y - \bar{Y})^2 = \Sigma Y^2 - n\bar{Y}^2 = 382 - 5\left(\frac{42}{5}\right)^2 = 382 - 352.8 = 29.2$$

To find a and b

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

or $42 = 5a + 16b$ .....(i)

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

or $160 = 16a + 74b$ .....(ii)

Coeff. of a	Coeff. of b	Constant
5	16	42
16	74	160

$$D = \begin{vmatrix} 5 & 16 \\ 16 & 74 \end{vmatrix} = 370 - 256 = 114$$

$$D_1 = \begin{vmatrix} 42 & 16 \\ 160 & 74 \end{vmatrix} = 3108 - 2560 = 548$$

$$D_2 = \begin{vmatrix} 5 & 42 \\ 16 & 160 \end{vmatrix} = 800 - 672 = 128$$

Now, $a = \frac{D_1}{D} = \frac{548}{114} = 4.807$

$$b = \frac{D_2}{D} = \frac{128}{114} = 1.122$$

$$SSE = \Sigma(Y - \hat{Y})^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY = 382 - 4.807 \times 42 - 1.122 \times 160$$
$$= 382 - 201.894 - 179.52 = 0.586$$

$$SSR = TSS - SSE = 29.2 - 0.586 = 28.614$$

Now $R^2 = \frac{SSR}{TSS} = \frac{28.614}{29.2} = 0.979$

EXERCISE

OBJECTIVE QUESTIONS

1. What is formula to obtain Karl Pearson's correlation coefficient?

   (a) $\frac{\text{Cov}(x, y)}{\Sigma d(x)}$ (b) $\frac{\text{Cov}(x, y)}{\Sigma d(y)}$ (c) $\frac{\text{Cov}(x, y)}{\Sigma d(x)\Sigma d(y)}$ (d) $\frac{\text{Cov}(x, y)}{V(x)\ v(y)}$

2. What is range of correlation coefficient?

   (a) -1 to 0 (b) 0 to 1 (c) - 1 to 1 (d) $-\infty$ to 1

3. When correlation coefficient between two variables is zero then what will be two variable

   (a) dependent (b) independent (c) increasing (d) decreasing

4. What does positive correlation implies?

   (a) as one variable is increasing then other variable is decreasing
   (b) as one variable is decreasing then other variable is increasing
   (c) as one variable is increasing then other variable is increasing
   (d) none

5. What is called functional relationship between two variables?

   (a) correlation (b) regression (c) skewness (d) kurtosis

6. What type of correlation exists between income and expenditure?

   (a) negative (b) no (c) positive (d) None

7. When inverse relationship exists between two variables then what is correlation between these variables?

   (a) Positive (b) Negative (c) No (d) Non sense

8. When direct relationship exists between two variables then what is correlation between these variables?

   (a) Positive (b) Negative (c) No (d) Non sense

9. When values of one variable decrease then values of other variable also decrease then what type of correlation exists?

   (a) Positive (b) Negative (c) Auto (d) No

10. When values of one variable increase then values of other variable decrease then what type of correlation exists?

    (a) Positive (b) Negative (c) Auto (d) No

11. What is relationship between correlation coefficient and regression coefficients?

    (a) $r = \sqrt{b_{yx} b_{xy}}$ (b) $r = \frac{b_{yx}}{\sigma_x}$ (c) $r = \frac{b_{xy}}{\sigma_y}$ (d) $r = \frac{b_{yx}}{\sigma_x} + \frac{b_{xy}}{\sigma_y}$

12. What is meaning of b in regression equation y = a+bx? (a) change in x per unit change in y (b) change in y per unit change in x (c) y intercept (d) dependent variable

13. What is dependent variable in regression equation y = a+bx? (a) x (b) y (c) a (d) b

14. What are equations by least square method to get a and b in regression equation y = a+bx?

(a) $\Sigma y = a + bx$, $\Sigma xy = ax + bx^2$ (b) $\Sigma y = a + b\Sigma x$, $\Sigma xy = a\Sigma x + b\Sigma x^2$

(c) $\Sigma y = na + b\Sigma x$, $\Sigma xy = a\Sigma x + b\Sigma x+2$ (d) $\Sigma y = a + bx$, $\Sigma xy = a\Sigma x + b\Sigma x^2$

15. The regression coefficient is independent of the change of

(a) Scale only (b) Origin only

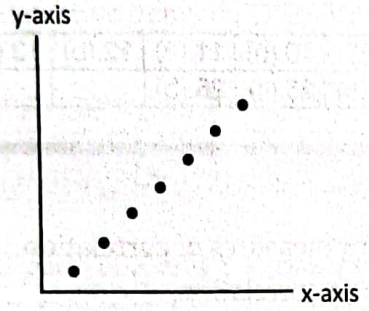
(c) Both scale and origin (d) Neither scale nor origin

16. The Karl Pearson's correlation coefficient is independent of change of

(a) origin (b) scale (c) origin and scale (d) all

17. What type of correlation exists if scatter plot is found as below?

(a) positive (b) perfect positive (c) negative (d) perfect negative



18. If r lies between 0.70 to 0.999 then what is type of correlation?

(a) high degree of positive (b) moderate degree of positive

(c) low degree of positive (d) none

19. If r lies between -0.699 to -0.500 then what is the type of correlation?

(a) high degree of negative (b) moderate degree of negative

(c) low degree of negative (d) none

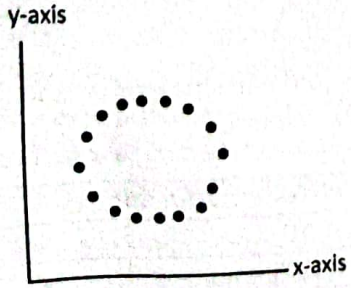
20. If r lies between -0.499 to -0.001 then what is the type of correlation?

(a) high degree of negative (b) moderate degree of negative

(c) low degree of negative (d) none

21. What type of correlation exists if scatter plot is found as below?

(a) positive (b) negative (c) non-sense (d) no



22. If covariance between x and y is 30, variance of x is 49 and variance of y is 100 then what is correlation coefficient between x and y?
(a) 3/7 (b) 3/10 (c) 30/49 (d) 7/10

23. In regression equation y = 0.8+1.23x what is meant by 1.23?
(a) y changes by 1.23 per unit change in x (b) y changes by 1 per 1.23 changes in x
(c) x changes by 1.23 per unit change in y (d) x changes by 1 per 1.23 change in y

24. The process of constructing a mathematical model that can be used to predict one variable using other variable is called
(a) correlation (b) regression (c) residual (d) outliers

25. What is the difference between actual value of dependent variable and estimated value of dependent variable obtained by using regression equation?
(a) correlation (b) regression (c) residual (d) intercept

26. Where does regression line of y on x and x on y intersect?
(a) (x,y) (b) $(\bar{x}, \bar{y})$ (c) (0,0) (d) none

ANSWERS

1.(c)	2.(c)	3.(b)	4.(c)	5.(b)	6.(c)	7.(b)	8.(a)	9.(a)	10.(b)	11.(a)	12.(b)	13.(b)	14.(c)	15.(b)
16.(d)	17.(b)	18.(a)	19.(b)	20.(c)	21.(d)	22.(a)	23.(a)	24.(b)	25.(c)	26.(b)				

THEORETICAL QUESTIONS

1. What do you mean by correlation? Describe different measures of correlation.
2. Distinguish between positive correlation and negative correlation.
3. Explain how Karl Pearson's correlation coefficient differs from Spearman's rank correlation coefficient?
4. Explain the suitable conditions for applying Karl Pearson's correlation coefficient and Spearman's rank correlation coefficient.
5. Write down properties of correlation coefficient.
6. What do you mean by significance of correlation? How can you interpret the correlation coefficient with the help of probable error?
7. What do you mean by regression? Interpret the regression coefficient.
8. What do you mean by regression coefficient? Write down its properties.
9. Differentiate between correlation and regression.
10. Describe the use of regression analysis in computer application.
11. How linear regression is different from nonlinear regression?
12. What do you mean by coefficient of determination? How can you interpret it?

NUMERICAL QUESTIONS

1. If the covariance between x and y variable is 36 and variance of X and Yare 36 and 100 respectively, find the coefficient of correlation between them. **[Ans: 0.6]**

2. Find the correlation coefficient between x and y series

	X	Y
Number of observation	10	10
Standard deviation	2.05	2.06

And $\Sigma(X - \bar{X})(Y - \bar{Y}) = 40$ **[Ans: 0.947]**

3. For 10 observations on two variables X and Y, the following information are as follows:
$\Sigma X = 666, \Sigma Y = 663, \Sigma X^2 = 44{,}490, \Sigma Y^2 = 44{,}061, \Sigma XY = 44{,}224$
Compute, Karl Pearson's coefficient of correlation. **[Ans: 0.576]**

4. From the following information find the total number of pair of observations given that $r = 0.8, \Sigma xy = 60$, S.D. of Y = 60 and $\Sigma x^2 = 90$, where x and y are deviations taken from their respective means. **[Ans: 1636]**

5. From the following data examine whether there exists any correlation between X and Y

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

[Ans: 0.95]

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

Size group	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45
No of items	200	270	340	360	400	300
No of defective items	150	162	170	180	180	120

[Ans: –0.939]

7. Following figure give the age in years of newly married husbands and wives (25,17) (26,18) (27,19) (25, 17) (26,19) (28,20) (25,17) (25,17) (24,18) (26,18) (26,20) (27,18) (27,19) (28,19) (25,18) (25,19) (26,18) (25,18) (27,20). Find Karl Pearson's Correlation coefficient. Test its significance. **[Ans: 0.654]**

8. From 20 pairs of X and Y variables the following results obtained
$\Sigma X = 127, \Sigma Y = 100, \Sigma X^2 = 860, \Sigma Y^2 = 549, \Sigma XY = 674$
at the time of verification, the following wrong values of X and Ywere taken as X = 10, 8 and Y= 14,6 instead of correct values X = 8,6 and Y= 12, 8. Find correct value of correlation. **[Ans: 0.47]**

9. A student calculates the value of r as 0.795 when the number of items is 100 and concludes that r is highly significant. Is his conclusion correct? **[Ans: Yes]**

10. Correlation coefficient between two variables with the pair of 10 observations is 0.81. Discuss if the value of r be significant or not. Also determine the limits of population correlation coefficient. **[Ans: Significant,0.736 to 0.883]**

11. The information given below are related with ages of husband(X) and wife(Y) for married couples living together in a sample survey. Calculate the coefficient of correlation between age of husband and that of his wife. Test the significance of the calculated r.
$N = 72, \Sigma fX = 3560, \Sigma fX^2 = 196800, \Sigma fY = 3260, \Sigma fY^2 = 168400, \Sigma fXY = 172000$
[Ans: 0.52,Significant]

12. A large company wants to measure the effectiveness of radio advertising media(x) on the sale promotion (y) of its products. A sample of 22 cities with approximately equal populations is selected for study. The sales of the product in thousand rupees and the level of radio advertising expenditure in thousand rupees are recorded for each of 22 cities. The sum, sum of square and sum of product of x and y are summarized below
$\sum y = 26953, \sum x = 950, \sum y^2 = 35528893, \sum x^2 = 49250, \sum xy = 1263940$
   a) Fit a simple linear regression model of y on x using the least square method. Interpret the estimated slope coefficient
   b) Compute R^2 and interpret. **[Ans: y=699.957+12.162x, 0.4852]**

13. The following measurements show the respective height in inches of 10 fathers and their eldest sons

Father	67	63	66	71	69	65	62	70	61	70
Son	68	66	65	70	67	67	64	71	62	68

Find the regression line of son's height on father's height and estimate the height of son for the given height of father as 70 inches. Also determine coefficient of determination and interpret. **[Ans: y = 40.43+0.388x, 67.62, 0.2...]**

14. The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Operator	I	II	III	IV	V	VI	VII	VIII
Experience	16	12	18	4	3	10	5	12
Performance	87	88	89	68	78	80	75	83

Calculate the regression equation of performance on experience and hence estimate the probable performance if an operator has 8 years experiences. Interpret the regression coefficient. **[Ans: y=69.669+1.133x,78.7...]**

15. A city council has gathered data on number of minor traffic accidents and the number of youth football games that occurred in town over the weekends.

X(football games)	20	30	10	12	15	25	34
Y(minor accidents)	6	9	4	5	7	8	9

(i) Develop the regression equation to predict minor accidents from football games.
(ii) Predict the number of minor traffic accidents that will occur at weekends during which X=30.
(iii) Calculate the value of coefficient of determination **[Ans: y=2.732+0.198x, 9, 0.87]**

16. A chemical company wishing to study the effect of extraction time on the efficiency of an extraction operation obtained the data as follows

Extraction time in minute(X)	27	45	41	19	35	39	19
Extraction efficiency in %(Y)	57	64	80	46	62	72	52

a) Fit a straight line to the given data by the method of least square and use it to predict the extraction efficiency one can expect when the extraction time is 35 minutes.
b) Determine the coefficient of determination and interpret its meaning.
[Ans: y=32.096+0.926x, 64.5, 0.843]

17. For the data given below i)Fit linear regression Y = a + bX by the method of least square and interpret regression coefficient ii) determine coefficient of determination and interpret.

X	0	5	10	15	20	25
Y	12	15	17	22	24	30

[Ans: y=11.29+0.697x, 0.97]

18. National Planning Commission (NPC) is performing preliminary study to determine the relationship between certain economic indicator and annual percentage change in Gross National Product (GNP). The concern is to estimate the percentage change in GNP. One of such indicator being examined is government's deficit. Data on 6 years are given below;

Percentage change in GNP	3	1	4	1	2	3
Government deficit in lakh Rs	50	200	70	100	90	40

a) Develop the estimating equation to predict percentage change in GNP from government deficit.
b) Interpret the estimated regression coefficient.
c) What percentage change in GNP would be expected in a year in which government deficit was Rs 110 lakh.
d) Compute the coefficient of determination and interpret.
[Ans: y = 3.725 – 0.015x, 2.055, 0.5241]

19. The annual advertising expenditure (in lakh Rs.) and the corresponding annual sales (in crore Rs.) for the past 10 years of a company are presented in the following table.

Year	Annual advertising expenditure	Annual sales revenue
1	10	20
2	12	30
3	14	37
4	16	50
5	18	56
6	20	78
7	22	89
8	24	100
9	26	120
10	28	110

a. Find the correlation coefficient between annual advertising expenditure and annual sales revenue and comment the result

b. Develop the regression model of sales as a function of advertising expenditures.

c. Predict the value of annual sales while advertising expenditure was 27 lakh rupees.

[Ans: 0.985,y=-40.048+5.739x,114.915]

20. Career airline pilots face the risk of progressive hearing loss due to the noisy cockpits of most jet aircrafts. Much of the noise comes not from engines but from air roar which increases at high speeds. To assess this workplace hazard a pilot measured cockpit noise level(in decibels) and airspeed (knots indicated air speed). The data are shown in the given table

Speed	250	340	320	330	346	260	280	395	380	400
Noise level	83	89	88	89	92	85	84	92	93	96

a. Determine association between noise level and air roar which is increased due to high speed. Comment on strength of association

b. Develop a least square regression model to estimate the noise level with the help of speed of aircraft. Also interpret the regression coefficient.

[Ans: 0.957;y=64.191+0.075x]

21. A computer manager interested to know how efficiency of his/ her new computer program which depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. Applying the program to data sets of different sizes, the following data were gathered.

Data size(gigabytes)	6	7	7	8	10	10	15
Processed requests	40	55	50	41	17	26	16

a. Identify which one response variable and fit a simple regression line assuming that the relationship is linear

b. Interpret the regression coefficient with reference to your problem

c. Obtain the coefficient of determination and interpret this

d. Based on the fitted model predict the efficiency of new computer for data size 12(gigabites). Does it possible to predict efficiency for data size of 30 (gigabites)? Discuss.

[Ans: y = 72.278- 4.142x, 0.661, 22.57, No.]

22. Omprakash Sharma, owner of the Kathmandu Precast Company, has hired you as a part-time analyst. He was extremely pleased when you uncovered a positive relationship between the number of building permits issued and the amount of work available to his company. Now he wonders if it's possible to use knowledge of interest rates on first mortgages to predict the number of building permits that will be issued each month. You collect a sample of data covering nine months.

Month	Building Permits (Y)	Interest rate (X) %
1	786	10.2
2	494	12.6
3	289	13.5
4	892	9.7
5	343	10.8
6	888	9.5
7	509	10.9
8	987	9.2
9	187	14.2

i. Calculate the correlation coefficient between building permits and interest rate and test its significance at 1%.
ii. Estimate the best fitting regression line and compute residual for month 9.
iii. Compute the coefficient of determination and interpret its meaning.
iv. Predict building permits when the interest rate increases by 9.7%.

[Ans: r = 0.89, Reject H_0, y = 2217.41 – 144.94x, $\hat{y}$ = 27.84, R^2 = 0.7933, $\hat{y}$ = 811.42]

23. Management of a soft-drink bottling company wants to develop a method for allocating delivery costs to customers. Although one cost clearly relates to travel time within a particular route, another variable cost reflects the time required to unload the cases of soft drink at the delivery point. A sample of 10 deliveries within a territory was selected. The delivery times and the number of cases delivered were recorded as follows:

Customer	Number of cases	Delivery times (minutes)
1	52	32.1
2	64	34.8
3	95	37.8
4	116	38.5
5	143	44.2
6	161	43.0
7	184	49.4
8	218	56.8
9	254	61.2
10	267	58.2

i. Find the correlation coefficient between delivery times and the number of cases delivered.
ii. Develop a regression model to predict delivery time, based on the number of cases delivered.
iii. Interpret the meaning of slope in this problem.
iv. Predict the delivery time for 150 cases of soft drink.
v. Determine the coefficient of determination and explain its meaning in this problem

[Ans: r = 0.109, y = 56.88 + 0.176x, $\hat{y}$ = 83.35, R^2 = 0.011]