MODULE - 4

STATISTICAL METHODS AND CURVE FITTING

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data. In other words, it is a Mathematical discipline to collect and summarize data.

Mathematical statistics is the application of Mathematics to Statistics, which was originally conceived as the science of the state — the collection and analysis of facts about a country: its economy, military, population and so forth.

Mathematical techniques used for this include mathematical analysis, linear algebra, stochastic analysis, differential equation and measure-theoretic probability theory.

Statistics is used in many sectors such as psychology, geology, sociology, weather forecasting, probability and much more. The goal of statistics is to gain understanding from data it focuses on applications and hence it is distinctively considered as a Mathematical science.

4.1 Basic Definitions and Formulae

Data: Data is a collection of facts, such as numbers, words, measurements, observations etc.

Classification: The significance of a large mass of statistical data known as raw data cannot be understood unless it is arranged in some definite manner. The process of arrangement in different groups is called classification.

Frequency table: It is tabular arrangement consisting of various classes of uniform size known as class intervals and the number in each class known as frequency. The average of the class interval in known as the midpoint of the class usually denoted by x.

Variate: It is a quantity which can have different numerical values.

Example: Price, Marks, Height, Weight, Age, …

Mean (Arithmetic mean): If $x_1, x_2, x_3, \dots, x_n$ are set of n values of a variate, then the mean is given by

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\Sigma x_i}{n}.$$

In a frequency distribution, if $x_1, x_2, x_3, \dots, x_n$ be the mid-values of the class-intervals having frequencies $f_1, f_2, f_3, \dots, f_n$ respectively, we have

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \dots + x_n f_n}{f_1 + f_2 + f_3 + \dots + f_n} = \frac{\Sigma x_i f_i}{\Sigma f_i}$$

Variance: If a variate x take the values $x_1, x_2, x_3, \dots, x_n$ then the variance V is defined as follows:

$$V = \frac{\Sigma\left(x - \bar{x}\right)^2}{n}$$

Also for a grouped data,

$$V = \frac{\Sigma f \left(x - \bar{x}\right)^2}{\Sigma f}$$

Standard deviation (S.D):

$$\pm = \sqrt{V} \text{ or } \pm^2 = V$$

$$\Rightarrow \qquad \pm = \sqrt{\frac{\Sigma\left(x - \bar{x}\right)^2}{n}} = \sqrt{\frac{\Sigma X^2}{n}} \text{ where } X = x - \bar{x}$$

$$\text{or} \qquad \pm^2 = \frac{\Sigma\left(x - \bar{x}\right)^2}{n} = \frac{\Sigma X^2}{n}$$

Alternative formula for $\pm^2$

$$\pm^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2$$

4.2. Correlation

The two variables *x* and *y* are related in such a way that an increase in one is accompanied by an increase or decrease in the other is called co-variation. Co-variation of two independent magnitudes is known as correlation.

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. Correlation is used to describe the linear relationship between two continuous variables. For example, height and weight are related; taller people tend to be heavier than shorter people.

Correlation is Positive when the values increase together and Correlation is Negative when one value decreases as the other increases.

4.2.1 Coefficient of correlation

The numerical measure of correlation between two variables *x* and *y* is known as Karl Pearson's coefficient of correlation or simply coefficient of correlation and is denoted by r and is defined by

$$r = \frac{\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{n\pm_x\pm_y} \qquad \text{----- (1)}$$

$$\Rightarrow \qquad r = \frac{\Sigma XY}{n\pm_x\pm_y} \qquad \text{----- (2)}$$

Where X = deviation from the mean = $x-\bar{x}$, Y = deviation from the mean = $y-\bar{y}$, n = number of values of the two variables, $\pm_x = \text{S.D. of } x-series$, $\pm_y = \text{S.D. of } y-series.$

Substituting the value of $\pm_x$ and $\pm_y$ in the above formula, we get

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}} \qquad \text{----- (3)}$$

Another form of the above formula is

$$r = \frac{n\Sigma xy - \Sigma x\Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\times\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}} \qquad \text{----- (4)}$$

Formula for correlation coefficient

The formula to compute coefficient of correlation is

$$r = \frac{\pm_x^2 + \pm_y^2 - \pm_{x-y}^2}{2\pm_x \pm_y}.$$

Proof:

Let $$z = x - y \qquad \text{----- (i)}$$

Taking summation and dividing by n, we get

$$\frac{\Sigma z}{n} = \frac{\Sigma x}{n} - \frac{\Sigma y}{n}$$

$\Rightarrow$ $$\bar{z} = \bar{x} - \bar{y} \qquad \text{----- (ii)}$$

Now, (i) – (ii) $\Rightarrow$ $$z - \bar{z} = (x - y) - (\bar{x} - \bar{y})$$

$\Rightarrow$ $$z - \bar{z} = (x - \bar{x}) - (y - \bar{y})$$

Squaring both sides, taking summation and dividing by n, we get

$$\frac{\sum (z-\bar{z})^2}{n} = \frac{\sum (x-\bar{x})^2}{n} + \frac{\sum (y-\bar{y})^2}{n} - \frac{2\sum (x-\bar{x})(y-\bar{y})}{n} \qquad \text{----- (iii)}$$

But, $$\pm^2 = \frac{\Sigma (x-\bar{x})^2}{n} \quad \text{and} \quad r = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\pm_x \pm_y}$$

Equation (iii) becomes

$$\pm_z^2 = \pm_x^2 + \pm_y^2 - 2r\pm_x \pm_y$$

$\Rightarrow$ $$\pm_{x-y}^2 = \pm_x^2 + \pm_y^2 - 2r\pm_x \pm_y$$

$\Rightarrow$ $$r = \frac{\pm_x^2 + \pm_y^2 - \pm_{x-y}^2}{2\pm_x \pm_y}$$

Property of coefficient of correlation

The coefficient of correlation numerically does not exceed '1',

i.e., $-1 \le r \le 1$.

4.3 Regression

It is an estimation of one independent variable in terms of the other.

The best fitting straight line of the form $y = a + bx$ is called the regression of y on x and $x = a + by$ is called the regression of x on y.

4.3.1 Equation of the Regression lines

Let $y = a + bx$ be the equation of the regression line of y on x for a given set on n values.

Its normal equations are

$$\Sigma y = na + b\Sigma x \qquad \text{----- (1)}$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \text{----- (2)}$$

Dividing (1) both side by 'n'

$$\frac{\Sigma y}{n} = a + \frac{b\Sigma x}{n}$$

$$\bar{y} = a + b\bar{x}$$

This shows that the regression line passes through $\left(\bar{x}, \bar{y}\right)$.

We know that the equation of a line in slope point form is

$$y - y_1 = m\left(x - x_1\right)$$

If $\left(x_1, y_1\right) = \left(\bar{x}, \bar{y}\right)$ then

$$y - \bar{y} = m\left(x - \bar{x}\right) \qquad \text{----- (3)}$$

$\Rightarrow$ $$Y = m\,X$$

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

The normal equation for $Y = m\,X$ is

$$\Sigma xy = m\Sigma x^2$$

$\Rightarrow$ $$m = \frac{\Sigma xy}{\Sigma x^2} \qquad \text{----- (4)}$$

We have, $$r = \frac{\Sigma xy}{n\pm_x\pm_y}$$

$\Rightarrow$ $$\Sigma xy = n\,r\pm_x\pm_y \qquad \text{----- (5)}$$

Also, we have, $$\pm_x^2 = \frac{\Sigma x^2}{n}$$

$$\Rightarrow \qquad \Sigma x^2 = n\pm_x^2 \qquad \text{----- (6)}$$

Using (5) & (6) in (4), we get

$$m = \frac{nr\pm_x \pm_y}{n\pm_x^2}$$

$$\Rightarrow \qquad m = \frac{r\pm_y}{\pm_x} \qquad \text{----- (7)}$$

Substituting (7) in (3), we get

$$y - \bar{y} = r\frac{\pm_y}{\pm_x}\left(x - \bar{x}\right)$$

This is the regression line of y on x.

Similarly, $x - \bar{x} = r\frac{\pm_x}{\pm_y}\left(y - \bar{y}\right)$ is the regression line of x on y

Here $r\frac{\pm_y}{\pm_x}$ and $r\frac{\pm_x}{\pm_y}$ are the regression coefficients. Their product is r^2

Thus 'r' is the geometric mean of regression coefficients,

$$\text{i.e., } r = \pm\sqrt{(\text{coefficient of } x)\times(\text{coefficient of } y)}$$

Note: If both the coefficients are positive then we have to consider r value as positive and if both coefficients are negative then we have to consider r value as negative.

4.3.2 Angle between the lines of regression

The angle between the lines of regression is

$$\tan^{-} = \frac{\pm_x \pm_y}{\pm_x^2 + \pm_y^2}\left(\frac{r^2 - 1}{r}\right)$$

Proof: **(VTU 2007, 2010, 2013, 2019)**

If m_1 and m_2 are the slopes of two lines then the angle between the lines is given by

----

$$\tan^{-} = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right| \qquad \text{----- (1)}$$

The lines of regression are

$$y - \bar{y} = r\frac{\pm_y}{\pm_x}\left(x - \bar{x}\right) \text{ and } x - \bar{x} = r\frac{\pm_x}{\pm_y}\left(y - \bar{y}\right)$$

$$\Rightarrow \qquad y - \bar{y} = r\frac{\pm_y}{\pm_x}\left(x - \bar{x}\right) \text{ and } y - \bar{y} = \frac{\pm_y}{r\pm_x}\left(x - \bar{x}\right)$$

The slopes of regression lines are $m_1 = \frac{r\pm_y}{\pm_x}$ and $m_2 = \frac{\pm_y}{r\pm_x}$

Using these in (1), we get

$$\tan^{-} = \left|\frac{\left(\frac{r\pm_y}{\pm_x}\right) - \left(\frac{\pm_y}{r\pm_x}\right)}{1 + \left(\frac{r\pm_y}{\pm_x}\right)\left(\frac{\pm_y}{r\pm_x}\right)}\right| = \left|\frac{\frac{\pm_y}{\pm_x}\left(r - \frac{1}{r}\right)}{1 + \frac{\pm_y^2}{\pm_x^2}}\right|$$

$$= \left|\frac{\frac{\pm_y}{\pm_x}\left(\frac{r^2 - 1}{r}\right)}{\left(\frac{\pm_x^2 + \pm_y^2}{\pm_x^2}\right)}\right| = \frac{\pm_x \pm_y}{\pm_x^2 + \pm_y^2}\left(\frac{r^2 - 1}{r}\right)$$

$\therefore$ The angle between the lines of regression is $\tan^{-} = \frac{\pm_x \pm_y}{\pm_x^2 + \pm_y^2}\left(\frac{r^2 - 1}{r}\right)$

Its significance when $r = 0$ and $r = \pm 1$

Case – 1: If $r = 0$ then $\tan^{-} = \infty$

$$\Rightarrow \qquad {}^{-} = \frac{®}{2}$$

$\therefore$ The lines are perpendicular

Case – 2: If $r = \pm 1$ then $\tan^{-} = 0$

$$\Rightarrow \qquad {}^{-} = 0$$

$\therefore$ The lines are parallel

WORKED EXAMPLES

Example 4.1: Find the coefficient of correlation and the lines of regression for the data

x	**1**	**2**	**3**	**4**	**5**
y	**2**	**5**	**3**	**8**	**7**

Solution: **(VTU 2006, 2008, 2010, 2016)**

We have, $$r=\frac{\pm_x^2+\pm_y^2-\pm_{x-y}^2}{2\pm_x\pm_y}$$

We prepare a relevant table as follows

	x	y	x J= y	x^2	y^2	$(x-y)^2$
	1	2	−1	1	4	1
	2	5	−3	4	25	9
	3	3	0	9	9	0
	4	8	−4	16	64	16
	5	7	−2	25	49	4
Σ	15	25	−10	55	151	30

Here, $n = 5$, $\Sigma x = 15$, $\Sigma y = 25$, $\Sigma(x$J=$y) = -10$, $\Sigma x^2 = 55$, $\Sigma y^2 = 151$ and $\Sigma(x-y)^2 = 30$.

$$\bar{x}=\frac{\Sigma x}{n}=\frac{15}{5}=3$$

$$\bar{y}=\frac{\Sigma y}{n}=\frac{25}{5}=5$$

$$\overline{(x-y)}=\frac{\Sigma(x-y)}{n}=-\frac{10}{5}=-2$$

$$\pm_x^2=\frac{\Sigma x^2}{n}-\left(\bar{x}\right)^2=\frac{55}{5}-(3)^2=2$$

$$\pm_y^2=\frac{\Sigma y^2}{n}-\left(\bar{y}\right)^2=\frac{151}{5}-(5)^2=5.2$$

$$\pm_{(x-y)^2}=\frac{\Sigma(x-y)^2}{n}-\left(\overline{x-y}\right)^2=\frac{30}{5}-(-2)^2=2$$

Substituting these values in r, we get

$$r=\frac{2+5.2-2}{2\left(\sqrt{2}\right)\left(\sqrt{5.2}\right)}=0.81$$

Also, we have the equations of the regression lines are

$$y-\bar{y}=r\frac{\pm_y}{\pm_x}\left(x-\bar{x}\right) \text{ and } x-\bar{x}=r\frac{\pm_x}{\pm_y}\left(y-\bar{y}\right)$$

$$y-5=(0.81)\frac{\sqrt{5.2}}{\sqrt{2}}(x-3) \text{ and } x-3=(0.81)\frac{\sqrt{2}}{\sqrt{5.2}}(y-5)$$

$$y-5=1.306(x-3) \text{ and } x-3=0.502(y-5)$$

$$y=1.306x+1.082 \text{ and } x=0.502y+0.49$$

These are the lines of regression.

Example 4.1.2: Find the coefficient of correlation and the lines of regression for the data

x	1	2	3	4	5	6	7
y	9	8	10	12	11	13	14

Solution: **(VTU 2007, 2010, 2013, 2018)**

We have, $$r=\frac{\pm_x^2+\pm_y^2-\pm_{x-y}^2}{2\pm_x\pm_y}$$

We prepare a relevant table as follows

	x	y	x J-y	x^2	y^2	$(x-y)^2$
	1	9	–8	1	81	64
	2	8	–6	4	64	36
	3	10	–7	9	100	49
	4	12	–8	16	144	64
	5	11	–6	25	121	36
	6	13	–7	36	169	49
	7	14	–7	49	196	49
Σ	28	77	–49	140	875	347

Here, n = 7, Σx = 28, Σy = 77, $\Sigma(x$J $y)$ = –49, Σx^2 = 140, Σy^2 = 875 and $\Sigma(x-y)^2$ = 347

$$\bar{x}=\frac{\Sigma x}{n}=\frac{28}{7}=4$$

$$\bar{y}=\frac{\Sigma y}{n}=\frac{77}{7}=11$$

$$\overline{(x-y)}=\frac{\Sigma(x-y)}{n}=-\frac{49}{7}=-7$$

$$\pm_x^2=\frac{\Sigma x^2}{n}-\left(\bar{x}\right)^2=\frac{140}{7}-(4)^2=4$$

$$\pm_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{875}{7} - (11)^2 = 4$$

$$\pm_{(x-y)^2} = \frac{\Sigma (x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{347}{7} - (-7)^2 = 0.57$$

Substituting these values in r, we get

$$r = \frac{4+4-0.57}{2(2)(2)} = 0.93$$

Also, we have the equations of the regression lines are

$$y - \bar{y} = r\frac{\pm_y}{\pm_x}\left(x - \bar{x}\right) \text{ and } x - \bar{x} = r\frac{\pm_x}{\pm_y}\left(y - \bar{y}\right)$$

$$y - 11 = (0.93)\frac{2}{2}(x-4) \text{ and } x - 4 = (0.93)\frac{2}{2}(y-11)$$

$$y - 11 = 0.93(x-4) \text{ and } x - 4 = 0.93(y-11)$$

$$y = 0.93x + 7.28 \text{ and } x = 0.93y - 6.23$$

These are the lines of regression.

Example 4.1.3: Find the coefficient of correlation and the lines of regression for the data

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

Solution: **(VTU 2005, 2007, 2010, 2016)**

We have, $$r = \frac{\pm_x^2 + \pm_y^2 - \pm_{x-y}^2}{2\pm_x \pm_y}$$

We prepare a relevant table as follows

	x	y	x J y	x^2	y^2	$(x-y)^2$
	1	9	–9	1	100	81
	2	12	–10	4	144	100
	3	16	–13	9	256	169
	4	28	–24	16	784	576
	5	25	–20	25	625	400
	6	36	–30	36	1296	900
	7	41	–34	49	1681	1156
	8	49	–41	64	2401	1681
	9	40	–31	81	1600	961
	10	50	–40	100	2500	1600
Σ	55	307	–252	385	11387	7624

Here, n = 10, Σx = 55, Σy = 307, $\Sigma(x$J $y)$ = –252, Σx^2 = 385, Σy^2 = 11387 and $\Sigma(x - y)^2$ = 7624

$$\bar{x} = \frac{\Sigma x}{n} = \frac{55}{10} = 5.5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{307}{10} = 30.7$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = -\frac{252}{10} = -25.2$$

$$\pm_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{385}{10} - (5.5)^2 = 8.25$$

$$\pm_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{11387}{10} - (30.7)^2 = 196.21$$

$$\pm_{(x-y)^2} = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{7624}{10} - (-25.2)^2 = 127.36$$

Substituting these values in r, we get

$$r = \frac{8.25 + 196.21 - 127.36}{2\left(\sqrt{8.25}\right)\left(\sqrt{196.21}\right)} = 0.96$$

Also, we have the equations of the regression lines are

$$y - \bar{y} = r\frac{\pm_y}{\pm_x}\left(x - \bar{x}\right) \text{ and } x - \bar{x} = r\frac{\pm_x}{\pm_y}\left(y - \bar{y}\right)$$

$$y - 30.7 = (0.96)\frac{\sqrt{196.21}}{\sqrt{8.25}}(x - 5.5) \text{ and } x - 5.5 = (0.96)\frac{\sqrt{8.25}}{\sqrt{196.21}}(y - 30.7)$$

$$y = 4.686x + 4.927 \text{ and } x = 0.197y - 0.548$$

These are the lines of regression.

Example 4.1.4: In a partially destroyed laboratory record of correlation data, the following results only are available:
Variance of x is 9 and the regression equations are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Calculate (i) the mean values of x and y, (ii) standard deviation of y and (iii) the coefficient of correlation between x and y.

Solution: **(VTU 2005, 2009, 2013)**

We know that the regression line passes through $\left(\bar{x}, \bar{y}\right)$

$$\therefore \quad 4\bar{x} - 5\bar{y} = -33$$
$$20\bar{x} - 9\bar{y} = 107$$

Solving these equations, we get

$$\bar{x} = 13 \text{ and } \bar{y} = 17$$

We rewrite the given regression equations as

$$y = \frac{4}{5}x + \frac{33}{5} \quad \text{and} \quad x = \frac{9}{20}y + \frac{107}{20} \qquad \text{----- (1)}$$

We have, the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$\Rightarrow \qquad r = \pm\sqrt{\left(\frac{4}{5}\right)\left(\frac{9}{20}\right)} = \pm\frac{3}{5}$$

$$\Rightarrow \qquad r = \frac{3}{5} \quad (\because \text{ both the coefficients are positive})$$

Given that, $\pm_x^2 = 9 \Rightarrow \pm_x = 3$

We have, $$y - \bar{y} = r\frac{\pm_y}{\pm_x}(x - \bar{x}) \qquad \text{----- (2)}$$

Comparing (2) with (1), we have $r\frac{\pm_y}{\pm_x} = \frac{4}{5}$

$$\Rightarrow \qquad \pm_y = \frac{4}{5} \times \frac{\pm_x}{r}$$

$$= \frac{4}{5} \times \frac{3}{(3/5)} = 4$$

Example 4.1.5: Compute $\bar{x}, \bar{y}$ and r from the equations of the regression lines $2x \text{ H } 3y \text{ H } 1 \text{ Z } 0$ and $x \text{ H } 6y \text{ Z } 4$.

Solution:

We know that the regression line passes through $(\bar{x}, \bar{y})$

$$\therefore \quad 2\bar{x} + 3\bar{y} = -1$$
$$\bar{x} + 6\bar{y} = 4$$

Solving these equations, we get

$$\bar{x} = -2 \text{ and } \bar{y} = 1$$

We rewrite the given regression equations as

$$y = -\frac{2}{3}x - \frac{1}{3} \quad \text{and} \quad x = -6y + 4 \qquad \text{----- (1)}$$

We have the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$\Rightarrow \qquad r = \pm\sqrt{\left(-\frac{2}{3}\right)(-6)} = \pm\sqrt{4} = \pm 2$$

which is not possible because $-1 \le r \le 1$.

Hence, we rewrite the given regression equations in other form as

$$x = -\frac{3}{2}y - \frac{1}{2} \quad \text{and} \quad y = -\frac{1}{6}x + \frac{2}{3} \qquad \text{----- (2)}$$

We have, the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$\Rightarrow \qquad r = \pm\sqrt{\left(-\frac{3}{2}\right)\left(-\frac{1}{6}\right)} = \pm\frac{1}{2}$$

$$\Rightarrow \qquad r = -\frac{1}{2} \quad (\because \text{both the coefficients are negative})$$

Example 4.1.6: If 2*x* J 3*y* = 0 and 3*x* J-2*y* = 5 are the lines of regression of the variables *x* and *y*. Find the following:

(i) Mean of *x* and *y*

(ii) Coefficient of correlation between *x* and *y*.

(iii) Angle between the lines of regression.

(iv) Standard deviation of *x* when variance of *y* is 2.

Solution:

(i) We know that the regression line passes through $(\bar{x}, \bar{y})$

$$\therefore \quad 2\bar{x} - 3\bar{y} = 0$$
$$3\bar{x} - 2\bar{y} = 5$$

Solving these equations, we get

$$\bar{x} = 3 \text{ and } \bar{y} = 2$$

(ii) We rewrite the given regression equations as

$$y = \frac{2}{3}x \quad \text{and} \quad x = \frac{2}{3}y + 5 \qquad \text{----- (1)}$$

We have the coefficient of correlation between x and y is

$$r = \pm\sqrt{(\text{coefficient of } x)(\text{coefficient of } y)}$$

$$\Rightarrow \qquad r = \pm\sqrt{\left(\frac{2}{3}\right)\left(\frac{2}{3}\right)} = \pm\frac{2}{3}$$

$$\Rightarrow \qquad r = \frac{2}{3} \quad (\because \text{both the coefficients are positive})$$

(iii) The angle between the lines of regression is $\tan{}^{-} = \left|\dfrac{m_1 - m_2}{1 + m_1 m_2}\right|$

where m_1 and m_2 are the slopes of two regression lines.

Now, $m_1 = -\dfrac{\text{coefficient of } x}{\text{coefficient of } y} = -\dfrac{2}{-3} = \dfrac{2}{3}$ and

$$m_2 = -\frac{\text{coefficient of } x}{\text{coefficient of } y} = -\frac{3}{-2} = \frac{3}{2}$$

$$\therefore \quad \tan{}^{-} = \left|\frac{(2/3) - (3/2)}{1 + (2/3)(3/2)}\right| = \left|\frac{-(5/6)}{2}\right| = \frac{5}{12}$$

$$\Rightarrow \qquad {}^{-} = \tan^{-1}\left(\frac{5}{12}\right)$$

(iv) Given that, $\pm_y^2 = 2 \Rightarrow \pm_y = \sqrt{2}$

We have, $y - \bar{y} = r\dfrac{\pm_y}{\pm_x}(x - \bar{x})$ ----- (2)

Comparing (2) with (1), we have $r\dfrac{\pm_y}{\pm_x} = \dfrac{2}{3}$

$$\Rightarrow \qquad \left(\frac{2}{3}\right)\frac{\sqrt{2}}{\pm_x} = \frac{2}{3}$$

$$\Rightarrow \qquad \pm_x = \sqrt{2}$$

Example 4.1.7: Find the lines of regression for the following data

Ages of cars (in years)	**2**	**4**	**6**	**7**	**8**	**10**	**12**
Annual Maintenance cost (in hundreds)	**16**	**15**	**18**	**19**	**17**	**21**	**20**

Hence estimate the maintenance cost if the age of a car is 9 years and the age of a car if the maintenance cost is Rs.1550/-

Solution:

We have, $$r = \frac{\pm_x^2 + \pm_y^2 - \pm_{x-y}^2}{2\pm_x \pm_y}$$

We prepare a relevant table as follows

	x	y	x J y	x^2	y^2	$(x-y)^2$
	2	16	–14	4	256	196
	4	15	–11	16	225	121
	6	18	–12	36	324	144
	7	19	–12	49	361	144
	8	17	–9	64	289	81
	10	21	–11	100	441	121
	12	20	–8	144	400	64
Σ	49	126	–77	413	2296	871

Here, $n = 7$, $\Sigma x = 49$, $\Sigma y = 126$, $\Sigma(x$J$y) = -77$, $\Sigma x^2 = 413$, $\Sigma y^2 = 2296$ and $\Sigma(x-y)^2 = 871$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{49}{7} = 7$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{7} = 18$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = -\frac{77}{7} = -11$$

$$\pm_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{413}{7} - (7)^2 = 10$$

$$\pm_y^2 = \frac{\Sigma y^2}{n} - \left(\bar{y}\right)^2 = \frac{2296}{7} - (18)^2 = 4$$

$$\pm_{(x-y)^2} = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{871}{7} - (-11)^2 = 3.429$$

Substituting these values in r, we get

$$r = \frac{10+4-3.429}{2\left(\sqrt{10}\right)\left(\sqrt{4}\right)} = 0.8357$$

Also, we have the equations of the regression lines are

$$y-\bar{y} = r\frac{\pm_y}{\pm_x}\left(x-\bar{x}\right) \text{ and } x-\bar{x} = r\frac{\pm_x}{\pm_y}\left(y-\bar{y}\right)$$

$$\Rightarrow \qquad y-18 = (0.8357)\frac{2}{\sqrt{10}}(x-7) \text{ and } x-7 = (0.8357)\frac{\sqrt{10}}{2}(y-18)$$

$$\Rightarrow \qquad y = 0.52857x+14.3 \text{ and } x = 1.321429y-16.786$$

These are the lines of regression.

Also,

(i) If the age of the car x = 9 then the maintenance cost y is given by the line of regression of y on x is

y = 0.52857 (9) + 14.3 = 19.06 hundreds.

$\therefore$ The maintenance cost = Rs.1906/-

(ii) If the maintenance cost y is Rs.1550/- i.e., 15.5 hundreds then the age of the car x is given by the line of regression of x on y is

x = 1.321429 (15.5) – 16.786 = 3.7 years

$\therefore$ Age of car = 3.7 years.

Example 4.1.8: In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales:

Salesmen	1	2	3	4	5	6	7	8	9	10
Test scores	40	70	50	60	80	50	90	40	60	60
Sales	2.5	6.0	4.5	5.0	4.5	2.0	5.5	3.0	4.5	3.0

Calculate the regression line of sales on test scores and estimate the most probable weekly sales volume if a salesman makes a score of 70.

Solution:

We have, $$r = \frac{\pm_x^2 + \pm_y^2 - \pm_{x-y}^2}{2\pm_x\pm_y}$$

We prepare a relevant table as follows

	Test scores x	Sales y	x J y	x^2	y^2	$(x-y)^2$
	40	2.5	37.5	1600	6.25	1406.25
	70	6.0	64	4900	36.0	4096
	50	4.5	45.5	2500	20.25	2070.25
	60	5.0	55	3600	25.0	3025
	80	4.5	75.5	6400	20.25	5700.25
	50	2.0	48	2500	4.0	2304
	90	5.5	84.5	8100	30.25	7140.25
	40	3.0	37	1600	9.0	1369
	60	4.5	55.5	3600	20.25	3080.25
	60	3.0	57	3600	9.0	3249
Σ	600	40.5	559.5	38400	180.25	33440.25

Here, $n = 10$, $\Sigma x = 600$, $\Sigma y = 40.5$, $\Sigma(x\text{J }y) = 559.5$, $\Sigma x^2 = 38400$, $\Sigma y^2 = 180.25$ and $\Sigma(x-y)^2 = 33440.25$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{600}{10} = 60$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{40.5}{10} = 4.05$$

$$\overline{(x-y)} = \frac{\Sigma(x-y)}{n} = \frac{559.5}{10} = 55.95$$

$$\pm_x^2 = \frac{\Sigma x^2}{n} - \left(\bar{x}\right)^2 = \frac{38400}{10} - (60)^2 = 240$$

$$\pm_y^2 = \frac{y^2}{n} - \left(\bar{y}\right)^2 = \frac{180.25}{10} - (4.05)^2 = 1.6225$$

$$\pm_{(x-y)^2} = \frac{\Sigma(x-y)^2}{n} - \left(\overline{x-y}\right)^2 = \frac{33440.25}{10} - (55.95)^2 = 213.6225$$

Substituting these values in r, we get

$$r = \frac{240 + 1.6225 - 213.6225}{2\left(\sqrt{240}\right)\left(\sqrt{1.6225}\right)} = 0.71$$

Also, we have the equations of the regression line of y on x is

$$y - \bar{y} = r\frac{\pm_y}{\pm_x}\left(x - \bar{x}\right)$$

$$\Rightarrow \qquad y - 4.05 = (0.71)\frac{\sqrt{1.6225}}{\sqrt{240}}(x - 60)$$

$$\Rightarrow \qquad y - 4.05 = 0.06(x - 60)$$

$$\Rightarrow \qquad y = 0.06x - 3.6 + 4.05$$

$$\Rightarrow \qquad y = 0.06x + 0.45$$

Now for x = 70 $\Rightarrow$ $\qquad y = 0.06(70) + 0.45 = 4.65$

Thus, the most probable weekly sales volume for a score 70 is 4.65

Example 4.1.9: While calculating correlation coefficient between two variables x and y from 25 pairs of observations, the following results were obtained: $n = 25$, $\Sigma x = 125$, $\Sigma y = 100$, $\Sigma x^2 = 650$, $\Sigma y^2 = 460$, $\Sigma xy = 508$. Later it was discovered at the time of checking that the pairs of values $\begin{array}{cc} x & y \\ 8 & 12 \\ 6 & 8 \end{array}$ were copied down as $\begin{array}{cc} x & y \\ 6 & 14 \\ 8 & 6 \end{array}$ Obtain the correct value of correlation coefficient.

Solution: **(VTU 2009, 2011, 2015)**

To get the correct results, we subtract the incorrect values and add the corresponding correct values.

$\therefore$ The correct results are

$$\begin{aligned}
&n = 25, \\
&\Sigma x = 125 - 6 - 8 + 8 + 6 = 125, \\
&\Sigma y = 100 - 14 - 6 + 12 + 8 = 100, \\
&\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650, \\
&\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436, \\
&\Sigma xy = 508 - 6\times 14 - 8\times 6 + 8\times 12 + 6\times 8 = 520
\end{aligned}$$

We have, $$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\left\{n\Sigma x^2 - (\Sigma x)^2\right\}\times\left\{n\Sigma y^2 - (\Sigma y)^2\right\}}}$$

$$\Rightarrow \qquad r = \frac{25\times 520 - 125\times 100}{\sqrt{\left\{25\times 650 - (125)^2\right\}\times\left\{25\times 436 - (100)^2\right\}}} = \frac{20}{\sqrt{25\times 36}} = \frac{2}{3}$$

EXERCISE 4.1

1. Find two lines of regression and coefficient of correlation for the data $n = 18, \Sigma x = 12, \Sigma y = 18, \Sigma x^2 = 6, \Sigma y^2 = 96, \Sigma xy = 48.$
2. Find the coefficient of correlation and the lines of regression for the data

x	2	4	6	8	10
y	5	7	9	8	11

3. Find the coefficient of correlation from the following data

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

4. Find the coefficient of correlation for the data

x	21	23	30	54	57	58	72	78	87	90
y	60	71	72	83	110	84	100	92	113	135

5. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{8}\right)$, show that $\pm_x = \frac{1}{2}\pm_y$. **(VTU 2004)**
6. Two random variables have the regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find the mean values and the correlation coefficient between x and y.
7. The regression equations of two variables x and y are $x = 0.7y + 5.2$ and $y = 0.3x + 2.8$. Find the means of the variables and the coefficient of correlation between them.
8. In a partially destroyed laboratory data, only the equations giving the two lines of regression of y on x and x on y are available and are respectively, $7x - 16y + 9 = 0$ and $5y - 4x - 3 = 0$. Calculate the coefficient of correlation and mean values of x and y.
9. The two regression equations of the variables x and y are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find the mean values of x and y and the correlation coefficient between x and y. **(VTU 2004)**
10. Psychological tests of intelligence and of engineering ability were applied to 10 students. Here is a record of ungrouped data showing intelligence ratio (I.R) and engineering ratio (E.R). Calculate the coefficient of correlation.

Student	A	B	C	D	E	F	G	H	I	J
I.R	105	104	102	101	100	99	98	96	93	92
E.R	101	103	100	98	95	96	104	92	97	94

ANSWERS

1. $r = 0.632$, $y = 0.467 + 0.8x$, $x = 0.167 + 0.5y$
2. $r = 0.91$, $x = 1.3y - 4.4$, $y = 0.65x + 4.1$
3. $r = 0.96$
4. 0.876
6. $\bar{x} = 4$, $\bar{y} = 7$, $r = -0.5$
7. $\bar{x} = 9.06$, $\bar{y} = 5.52$, $r = 0.46$
8. $r = 0.7395$, $\bar{x} = -0.1034$, $\bar{y} = 0.5172$
9. $\bar{x} = 9.0$ or 15.79, $\bar{y} = 3.74$, $r = -0.66$
10. $r = 0.59$

4.4 Rank Correlation

A group of individuals may be arranged in order to merit with respect to some characteristic. The same group would give different orders for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the rank correlation in the characteristics A and B for that group of individuals.

Let x_i and y_i be the ranks of the i^{th} individuals in A and B respectively. Assuming that no two individuals are bracketed equal in either case, each of the variables taking the values 1, 2, … , n,

we have, $$\bar{x} = \bar{y} = \frac{1+2+3+\ldots+n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

If X and Y be the deviations of x and y from their means, then

$$\Sigma X_i^2 = \Sigma\left(x_i - \bar{x}\right)^2 = \Sigma x_i^2 + n\left(\bar{x}\right)^2 - 2\bar{x}\Sigma x_i$$

$$= \Sigma n^2 + \frac{n(n+1)^2}{4} - 2\left(\frac{n+1}{2}\right)\Sigma n$$

$$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)^2}{4} - \frac{n(n+1)^2}{2} = \frac{1}{12}\left(n^3 - n\right)$$

Similarly, $\Sigma Y_i^2 = \frac{1}{12}\left(n^3 - n\right)$

Now, let $d_i = x_i - y_i$ so that $d_i = \left(x_i - \bar{x}\right) - \left(y_i - \bar{y}\right) = X_i - Y_i$

$$\therefore \qquad \Sigma d_i^2 = \Sigma X_i^2 + \Sigma Y_i^2 - 2\Sigma X_i Y_i$$

$$\Rightarrow \qquad \Sigma X_i Y_i = \frac{1}{2}\left[\Sigma X_i^2 + \Sigma Y_i^2 - \Sigma d_i^2\right]$$

$$\Rightarrow \qquad \Sigma X_i Y_i = \frac{1}{12}\left(n^3 - n\right) - \frac{1}{2}\Sigma d_i^2$$

Hence the correlation coefficient between these variables is

$$r = \frac{\Sigma X_i Y_i}{\sqrt{\Sigma X_i^2 \Sigma Y_i^2}} = \frac{\frac{1}{12}\left(n^3 - n\right) - \frac{1}{2}\Sigma d_i^2}{\frac{1}{12}\left(n^3 - n\right)}$$

$$\Rightarrow \qquad r = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

This formula is called the rank correlation coefficient or Spearman's Rank Correlation Coefficient and is denoted by °.

$$\text{i.e.,} \quad ° = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

Spearman's rank is probably one of the most useful statistical tests that we can do in Geography to prove a relationship between two different sets of data.

The Spearman rank correlation coefficient °, is the non-parametric version of the Pearson correlation coefficient. Data must be ordinal, interval or ratio. Spearman's returns a value from –1 to 1, where:

+1 = a perfect positive correlation between ranks

–1 = a perfect negative correlation between ranks

0 = no correlation between ranks.

Note: To assign the rank for the given set of values, order the scores from greatest to smallest; assign the rank 1 to the highest score, 2 to the next highest and so on. If ranks are tied, i.e., Tied ranks are where two items in a column have the same rank then each tied data point assigned a mean rank.

WORKED EXAMPLES

Example 4.4.1: Calculate the rank correlation coefficient and comment on the following data on sunflower:

Height of a sunflower (in cm)	**183**	**134**	**234**	**256**	**190**	**89**	**112**
Width of the stem (in mm)	**21**	**14**	**24**	**32**	**29**	**18**	**20**

Solution:

We prepare a relevant table as follows

Height of a sunflower (in cm)	Rank x_i	Width of the stem (in mm)	Rank y_i	$d_i = x_i - y_i$	d_i^2
183	4	21	4	0	0
134	5	14	7	-2	4
234	2	24	3	-1	1
256	1	32	1	0	0
190	3	29	2	1	1
89	7	18	6	1	1
112	6	20	5	1	1
				Σ	8

Here, $n = 7$, $\Sigma d_i^2 = 8$

We have, the rank correlation coefficient is

$$° = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

$$\Rightarrow \quad ° = 1 - \frac{6(8)}{\left(7^3 - 7\right)} = 1 - 0.143 = 0.857$$

As ° is close to one, we can conclude that the wider the stem to higher the sunflower grows.

Example 4.4.2: The scores for 9 students in Physics and Maths are as follows:

Physics:	**35**	**23**	**47**	**17**	**10**	**43**	**9**	**6**	**28**
Mathematics:	**30**	**33**	**45**	**23**	**8**	**49**	**12**	**4**	**31**

Compute the ranks of students in the two subjects and compute the Spearman's rank correlation.

Solution:

We prepare a relevant table as follows

Physics	Rank x_i	Mathematics	Rank y_i	$d_i = x_i - y_i$	d_i^2
35	3	30	5	-2	4
23	5	33	3	2	4
47	1	45	2	-1	1
17	6	23	6	0	0
10	7	8	8	-1	1
43	2	49	1	1	1
9	8	12	7	1	1
6	9	4	9	0	0
28	4	31	4	0	0
				Σ	12

Here, $n = 9$, $\Sigma d_i^2 = 12$

We have, the rank correlation coefficient is

$$° = 1 - \frac{6\Sigma d_i^2}{\left(n^3 - n\right)}$$

$$\Rightarrow \quad ° = 1 - \frac{6(12)}{\left(9^3 - 9\right)} = 1 - 0.1 = 0.9$$

Example 4.4.3: Ten participants in a contest are ranked by two judges as follows:

X	**1**	**6**	**5**	**10**	**3**	**2**	**4**	**9**	**7**	**8**
Y	**6**	**4**	**9**	**8**	**1**	**2**	**3**	**10**	**5**	**7**

Calculate the rank correlation coefficient.

Solution: **(VTU 2002)**

We prepare a relevant table as follows

Rank x_i	Rank y_i	$d_i = x_i - y_i$	d_i^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
		Σ	60

Here, n = 10, $\Sigma d_i^2 = 60$

We have, the rank correlation coefficient is

$$° = 1 - \frac{6\Sigma d_i^2}{(n^3 - n)}$$

$$\Rightarrow \quad ° = 1 - \frac{6(60)}{(10^3 - 10)} = 1 - 0.36 = 0.64$$

Example 4.4.4: Three judges A, B and C give the following ranks. Find which pair of judges has common approach

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

Solution: **(VTU 2003)**

We prepare a relevant table as follows

Ranks by A x_i	Ranks by B y_i	Ranks by C z_i	d_1 $x_i - y_i$	d_2 $y_i - z_i$	d_3 $z_i - x_i$	d_1^2	d_2^2	d_2^2
1	3	6	-2	-3	5	4	9	25
6	5	4	1	1	-2	1	1	4
5	8	9	-3	-1	4	9	1	16
10	4	8	6	-4	-2	36	16	4
3	7	1	-4	6	-2	16	36	4
2	10	2	-8	8	0	64	64	0
4	2	3	2	-1	-1	4	1	1
9	1	10	8	-9	1	64	81	1
7	6	5	1	1	-2	1	1	4
8	9	7	-1	2	1	1	4	1
					Σ	200	214	60

Here, $n = 10$

We have, the rank correlation coefficient is

$$\circ = 1 - \frac{6\Sigma d_i^{\,2}}{(n^3 - n)}$$

$$\therefore \qquad \circ(x, y) = 1 - \frac{6\Sigma d_1^{\,2}}{(n^3 - n)} = 1 - \frac{6(200)}{(10^3 - 10)} = -0.2$$

$$\circ(y, z) = 1 - \frac{6\Sigma d_2^{\,2}}{(n^3 - n)} = 1 - \frac{6(214)}{(10^3 - 10)} = -0.3$$

$$\circ(z, x) = 1 - \frac{6\Sigma d_3^{\,2}}{(n^3 - n)} = 1 - \frac{6(60)}{(10^3 - 10)} = 0.6$$

Since $\circ(z, x)$ is maximum, the pair of judges A and C have the nearest common approach.

EXERCISE 4.2

1. Find the rank correlation for the following data:

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

2. Calculate the rank correlation coefficient from the following data showing ranks of 10 students in two subjects:

Maths:	3	8	9	2	7	10	4	6	1	5
Physics:	5	9	10	1	8	7	3	4	2	6

3. Find the rank correlation coefficient for the following data

x	68	64	75	50	64	80	75	40	55	64
y	62	58	68	45	81	60	68	48	50	70

(VTU 2018)

ANSWERS

1. 0.932

2. 0.8545

4.5 Curve Fitting

It is the method of finding equation of a curve that approximates a given set of *n* data points and this equation is called best fitting equation.

4.5.1 Method of Least squares

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a given set of *n* data points. Here y_i's are called the observed values and $Y_i = f(x_i)$ are called the estimated values. The difference $d_i = y_i - Y_i$ are called residuals or errors. The method of least squares provides a relationship $y = f(x)$ such that sum of the squares of the residuals is minimum, i.e., $\sum_{i=1}^{n} d_i^2$ is minimum.

4.5.2 Fitting of a straight line *y* = *ax* + *b*

Consider a set of '*n*' given values of *x* and *y* for fitting a straight line

$$y = ax + b \qquad \text{----- (1)}$$

The residual

$$d = y - (ax + b) \qquad \text{----- (2)}$$

is the difference between the observed and estimated values of y.

Let $$S = \sum d^2$$

$\Rightarrow$ $$S = \sum \left[y - (ax + b) \right]^2 \qquad \text{----- (3)}$$

We determine *a* and *b* so that *S* is minimum.

The necessary conditions for this are

$$\frac{\partial S}{\partial a} = 0 \text{ and } \frac{\partial S}{\partial b} = 0 \qquad \text{----- (4)}$$

From (3), we find

$$\frac{\partial S}{\partial a} = \sum 2\left[y - (ax + b) \right](-x) \text{ and } \frac{\partial S}{\partial b} = \sum 2\left[y - (ax + b) \right](-1)$$

$\Rightarrow$ $$\frac{\partial S}{\partial a} = -2\sum \left[y - (ax + b) \right] x \text{ and } \frac{\partial S}{\partial b} = -2\sum \left[y - (ax + b) \right]$$

Using these in (4), we get

$$-2\sum\left[y-(ax+b)\right]x=0 \text{ and } -2\sum\left[y-(ax+b)\right]=0$$

$$\Rightarrow \quad \sum\left[y-(ax+b)\right]x=0 \text{ and } \sum\left[y-(ax+b)\right]=0$$

$$\Rightarrow \quad \sum_1^n xy-\sum_1^n ax^2-\sum_1^n bx=0 \text{ and } \sum_1^n y-\sum_1^n ax-\sum_1^n b=0$$

But $\sum_1^n b = b+b+b+\ldots+b = nb$

$$\therefore \quad \Sigma y = a\Sigma x + nb \quad \text{and}$$
$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

These equations are called normal equations for fitting a straight line $y = ax + b$

Note: The normal equations for fitting a straight line $y = a + bx$ are

$$\Sigma y = na + b\Sigma x \quad \text{and}$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

Working procedure to fit a straight line $y = ax + b$ *or* $y = a + bx$

- Write the normal equations of the given curve.
- Prepare the relevant table and find the value of summation present in the normal equation and substitute.
- We find the parameters '*a*' and '*b*' by solving and then substitute in the given equation.

WORKED EXAMPLES

Example 4.5.1: Fit a straight line $y = ax + b$ for the data

x	0	1	2	3	4	5	6
y	J:4	J2	0	2	4	6	8

Solution:

The normal equations of fitting $y = ax + b$ are

$$\therefore \quad \Sigma y = a\Sigma x + nb \quad \text{and}$$
$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

We prepare a relevant table as follows

	x	y	xy	x^2
	0	– 4	0	0
	1	–2	–2	1
	2	0	0	4
	3	2	6	9
	4	4	16	16
	5	6	30	25
	6	8	48	36
Σ	21	14	98	91

Here, $n = 7$, $\Sigma x = 21$, $\Sigma y = 14$, $\Sigma xy = 98$ and $\Sigma x^2 = 91$

Substituting these values in the above normal equations, we get

$$14 = 21a + 7b$$
$$98 = 91a + 21b$$

Solving these equations, we get

$$a = 2, \quad b = -4$$

$\therefore$ The equation of the best fitting straight line is $y = 2x - 4$

Example 4.5.2: Fit a straight line $y = ax + b$ for the data

x	**5**	**10**	**15**	**20**	**25**
y	**16**	**19**	**23**	**26**	**30**

Solution:

The normal equations of fitting $y = ax + b$ are

$$\therefore \quad \Sigma y = a\Sigma x + nb \quad \text{and}$$
$$\Sigma xy = a\Sigma x^2 + b\Sigma x$$

We prepare a relevant table as follows

	x	y	xy	x^2
	5	16	80	25
	10	19	190	100
	15	23	345	225
	20	26	520	400
	25	30	750	625
Σ	75	114	1885	1375

Here, n = 5, Σx = 75, Σy = 114, Σxy = 1885 and Σx^2 = 1375

Substituting these values in the above normal equations, we get

$$114 = 75a + 5b$$
$$98 = 1375a + 75b$$

Solving these equations, we get

$$a = 0.7, \quad b = 12.3$$

$\therefore$ The equation of the best fitting straight line is $y = 0.7x + 12.3$

Example 4.5.3: Fit a straight line $y = a + bx$ for the data

x	0	1	3	6	8
y	1	3	2	5	4

Solution:

The normal equations of fitting $y = ax + b$ are

$$\Sigma y = na + b\Sigma x \text{ and}$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

We prepare a relevant table as follows

	x	y	xy	x^2
	0	1	0	0
	1	3	3	1
	3	2	6	9
	6	5	30	36
	8	4	32	64
Σ	18	15	71	110

Here, n = 5, Σx = 18, Σy = 15, Σxy = 71 and Σx^2 = 110

Substituting these values in the above normal equations, we get

$$15 = 5a + 18b$$
$$71 = 18a + 110b$$

Solving these equations, we get

$$a = 1.6, \quad b = 0.38$$

$\therefore$ The equation of the best fitting straight line is $y = 1.6 + 0.38x$

Example 4.5.4: If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form $P = mW + c$ connecting P and W, using the following data:

P	**12**	**15**	**21**	**25**
W	**50**	**70**	**100**	**120**

Solution: **(VTU 2007, 2013)**

The normal equations of fitting $P = mW + c$ are

$$\Sigma P = m\Sigma W + nc \quad \text{and}$$
$$\Sigma WP = m\Sigma W^2 + c\Sigma x$$

We prepare a relevant table as follows

	W	P	WP	W^2
	50	12	600	2500
	70	15	1050	4900
	100	21	2100	10000
	120	25	3000	14400
Σ	340	73	6750	31800

Here, $n = 4$, $\Sigma W = 340$, $\Sigma P = 73$, $\Sigma WP = 6750$ and $\Sigma W^2 = 31800$

Substituting these values in the above normal equations, we get

$$73 = 340m + 4c$$
$$6750 = 31800m + 340c$$

Solving these equations, we get

$$m = 0.1879, \quad c = 2.2759$$

$\therefore$ The best fit of a line is $P = 0.1879W + 2.2759$

When W = 150 kg $\Rightarrow P = 0.1879(150) + 2.2759 = 30.4635$ kg

EXERCISE 4.3

1. Fit a straight line to the following data

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	5

2. Fit a straight line to the following data

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

(VTU 2011)

3. Fit a straight line to the following data

Year x	1961	1971	1981	1991	2001
Production (in tons) y	8	10	12	10	16

and find the expected production in 2006.

4. A simply supported beam carries a concentrated load P at its mid-point. Corresponding to various values of P, the maximum deflection Y is measured. The data are given below:

P	100	120	140	160	180	200
y	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $Y = a + bP$

5. The results of measurement of electric resistance R of a copper bar at various temperatures $t\,^0C$ are listed below:

t	19	25	30	36	40	45	50
R	76	77	79	80	82	83	85

Find a relation $R = a + bt$ where a and b are constants to be determined.

6. Fit a straight line to the following data

x	0	0.2	0.5	0.7	0.9	1.1	1.3
y	1.5	1.08	0.45	0.03	–0.39	–0.81	–1.23

ANSWERS

1. $y = 4.193 + 1.117x$
2. $y = 0.5454 + 0.6363x$
3. 15.2
4. $Y = 0.004P + 0.048$
5. $R = 70.052 + 0.292t$
6. $y = 1.5 - 2.1x$

4.5.3 Fitting of a curve of the form $y = ax^b$

Consider a set of 'n' given values of x and y for fitting a Geometric curve

$$y = ax^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log y = \log\left(ax^b\right)$$

$$\log y = \log a + b\log x$$

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log y$, $A = \log a$, $B = b$ and $X = \log x$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

Solving these equations we obtain A and B from which $a = e^A$ and $b = B$ can be found.

Substituting these values of a and b in (1), we obtain the equation of the best fitting curve.

WORKED EXAMPLES

Example 4.5.5: Fit a curve of the form $y = ax^b$ for the data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.1	6.8	7.5

Solution:

$$y = ax^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log y = \log\left(ax^b\right)$$

$$\log y = \log a + b\log x$$

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log y$, $A = \log a$, $B = b$ and $X = \log x$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$
$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

We prepare a relevant table as follows

	x	y	$X = \log x$	$Y = \log y$	XY	X^2
	1	2.98	0	1.0919	0	0
	2	4.26	0.6931	1.4492	1.0044	0.4804
	3	5.21	1.0986	1.6506	1.8133	1.2069
	4	6.1	1.3863	1.8083	2.5068	1.9218
	5	6.8	1.6094	1.9169	3.0851	2.5909
	6	7.5	1.7918	2.0149	3.6103	3.2105
Σ	21	32.85	6.5792	9.9318	12.0199	9.4098

Here, $n = 6$, $\Sigma X = 6.5792$, $\Sigma Y = 9.9318$, $\Sigma XY = 12.0199$ and $\Sigma X^2 = 9.4098$

Substituting these values in the above normal equations, we get

$$9.9318 = 6A + 6.5792B$$
$$12.0199 = 6.5792A + 9.4098B$$

Solving these equations, we get

$$A = 1.0912, \quad B = 0.5144$$

$$\Rightarrow \quad a = e^A = e^{1.0912} = 2.9778 \text{ and } b = B = 0.5144$$

$\therefore$ The equation of the best fitting curve is $y = (2.9778)x^{0.5144}$

Example 4.5.6: Fit a curve of the form $y = ax^b$ for the data

x	**1**	**2**	**3**	**4**	**5**
y	**0.5**	**2**	**4.5**	**8**	**12.5**

Solution:

$$y = ax^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log y = \log\left(ax^b\right)$$
$$\log y = \log a + b\log x$$

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log y,\ A = \log a,\ B = b$ and $X = \log x$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$
$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

We prepare a relevant table as follows

	x	y	$X = \log x$	$Y = \log y$	XY	X^2
	1	0.5	0	-0.6931	0	0
	2	2	0.6931	0.6931	0.4804	0.4804
	3	4.5	1.0986	1.5041	1.6524	1.2069
	4	8	1.3863	2.0794	2.8827	1.9218
	5	12.5	1.6094	2.5257	4.0649	2.5903
Σ	15	27.5	4.7874	6.1092	9.0804	6.1993

Here, $n = 5$, $\Sigma X = 4.7874$, $\Sigma Y = 6.1092$, $\Sigma XY = 9.0804$ and $\Sigma X^2 = 6.1993$

Substituting these values in the above normal equations, we get

$$6.1092 = 5A + 4.7874B$$
$$9.0804 = 4.7874A + 6.1993B$$

Solving these equations, we get

$$A = -0.6929, \quad B = 1.9998$$

$$\Rightarrow \quad a = e^A = e^{-0.6929} = 0.5 \quad \text{and} \quad b = B = 1.9998$$

$\therefore$ The equation of the best fitting curve is $y = (0.5)x^{1.9998}$

Example 4.5.7: An experiment gave the following values:

***v* (ft/min)**	**350**	**400**	**500**	**600**
***t* (min)**	**61**	**26**	**7**	**2.6**

It is known that *v* and *t* are connected by the relation $v = at^b$. Find the best possible values of *a* and *b*.

Solution: **(VTU 2013)**

$$v = at^b \qquad \text{----- (1)}$$

Taking *log* on both sides

$$\log v = \log\left(a t^b\right)$$
$$\log v = \log a + b \log t$$

$$Y = A + BX \qquad \text{----- (2)}$$

where $Y = \log v,\ A = \log a,\ B = b$ and $X = \log t$

The normal equations of (2) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and}$$
$$\Sigma XY = A\Sigma X + B\Sigma X^2$$

We prepare a relevant table as follows

	v	t	$X = \log t$	$Y = \log v$	XY	X^2
	350	61	4.1109	5.8579	24.0812	16.8995
	400	26	3.2581	5.9915	19.5209	10.6152
	500	7	1.9459	6.2146	12.0930	3.7865
	600	2.6	0.9555	6.3969	6.1122	0.9130
Σ	1850	96.6	10.2704	24.4609	61.8073	29.2142

Here, $n = 4$, $\Sigma X = 10.2704$, $\Sigma Y = 24.4609$, $\Sigma XY = 61.8073$ and $\Sigma X^2 = 29.2142$

Substituting these values in the above normal equations, we get

$$24.4609 = 4A + 10.2704B$$
$$61.8073 = 10.2704A + 29.2142B$$

Solving these equations, we get

$$A = 7.0167, \quad B = -0.3511$$

$$\Rightarrow \quad a = e^A = e^{7.0167} = 1115.10 \quad \text{and} \quad b = B = -0.3511$$

$\therefore$ The equation of the best fitting curve is $y = (0.5)x^{1.9998}$

EXERCISE 4.4

1. Fit a curve of the form $y = ax^b$ for the data

x	20	16	10	11	14
y	22	41	120	89	56

(VTU 2015)

2. Fit a curve of the form $y = ax^b$ for the data

x	1	2	4	6
y	6	4	2	2

3. Predict y at x = 3.75, by fitting a curve $y = ax^b$ for the data

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.10	6.80	7.50

4. Fit a curve of the form $y = ax^b$ for the data

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

ANSWERS

1. $y = (28491)(x)^{-2.38}$

2. $y = (5.965)(x)^{-0.672}$

3. $y = (2.978)(x)^{0.5143}$; 5.8769

4. $y = (0.5)x^2$

4.5.4 Fitting of a Second degree curve (Parabola) y Z ax^2 H bx H c

Consider a set of 'n' given values of x and y for fitting a second degree curve

$$y = ax^2 + bx + c \qquad \text{----- (1)}$$

The residual

$$d = y - \left(ax^2 + bx + c\right) \qquad \text{----- (2)}$$

is the difference between the observed and estimated values of y.

Let
$$S = \sum d^2$$

$\Rightarrow$
$$S = \sum\left[y - \left(ax^2 + bx + c\right)\right]^2 \qquad \text{----- (3)}$$

We determine a, b and c so that S is minimum.

The necessary conditions for this are

$$\frac{\partial S}{\partial a} = 0, \quad \frac{\partial S}{\partial b} = 0 \text{ and } \frac{\partial S}{\partial c} = 0 \qquad \text{----- (4)}$$

From (3), we find

$$\frac{\partial S}{\partial a} = \sum 2\left[y - \left(ax^2 + bx + c\right)\right]\left(-x^2\right), \quad \frac{\partial S}{\partial b} = \sum 2\left[y - \left(ax^2 + bx + c\right)\right](-x)$$

$$\text{and} \quad \frac{\partial S}{\partial c} = \sum 2\left[y - \left(ax^2 + bx + c\right)\right](-1)$$

$$\frac{\partial S}{\partial a}=-2\sum\left[y-\left(ax^2+bx+c\right)\right]\left(x^2\right),\ \frac{\partial S}{\partial b}=-2\sum\left[y-\left(ax^2+bx+c\right)\right](x)$$

$$\text{and}\quad \frac{\partial S}{\partial c}=-2\sum\left[y-\left(ax^2+bx+c\right)\right]$$

Using these in (4), we get

$$-2\sum\left[y-\left(ax^2+bx+c\right)\right]\left(x^2\right)=0,\ -2\sum\left[y-\left(ax^2+bx+c\right)\right](x)=0$$

$$\text{and}\quad -2\sum\left[y-\left(ax^2+bx+c\right)\right]=0$$

$$\Rightarrow\sum\left[y-\left(ax^2+bx+c\right)\right]\left(x^2\right)=0,\ \sum\left[y-\left(ax^2+bx+c\right)\right](x)=0$$

$$\text{and}\quad \sum\left[y-\left(ax^2+bx+c\right)\right]=0$$

$$\Rightarrow\ \sum_1^n x^2y-\sum_1^n ax^4-\sum_1^n bx^3-\sum_1^n cx^2=0,\ \sum_1^n xy-\sum_1^n ax^3-\sum_1^n bx^2-\sum_1^n cx=0$$

$$\text{and}\quad \sum_1^n y-\sum_1^n ax^2-\sum_1^n bx-\sum_1^n c=0$$

But $\sum_1^n c=c+c+c+...+c=nc$

$$\therefore\quad \Sigma y=a\Sigma x^2+b\Sigma x+nc$$
$$\Sigma xy=a\Sigma x^3+b\Sigma x^2+c\Sigma x\quad \text{and}$$
$$\Sigma x^2y=a\Sigma x^4+b\Sigma x^3+c\Sigma x^2$$

These equations are called normal equations for fitting a second degree curve $y=ax^2+bx+c$

Note: The normal equations for fitting a straight line $y=a+bx+cx^2$ are

$$\Sigma y=na+b\Sigma x+c\Sigma x^2$$
$$\Sigma xy=a\Sigma x+b\Sigma x^2+c\Sigma x^3\quad \text{and}$$
$$\Sigma x^2y=a\Sigma x^2+b\Sigma x^3+c\Sigma x^4$$

Working procedure to fit a second degree curve $y = ax^2 + bx + c$ or $y = a + bx + cx^2$

- ☞ Write the normal equations of the curve
- ☞ Prepare the relevant table and find the value of summation present in the normal equation and substitute.
- ☞ We find the parameters '*a*' '*b*' and '*c*' by solving and then substitute in the given equation.

WORKED EXAMPLES

Example 4.5.8: Fit a second degree curve $y = ax^2 + bx + c$ for the data

x	1.0	1.5	2.0	2.5	3.0	3.5	4.0
y	NKN=	NKP=	1.6	2.0	2.7	3.4	4.1

Solution: **(VTU 2004, 2005, 2009)**

The normal equations of fitting $y = ax^2 + bx + c$ are

$$\therefore \quad \Sigma y = a\Sigma x^2 + b\Sigma x + nc$$
$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{and}$$
$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	1.0	1.1	1.1	1	1	1	1.1
	1.5	1.3	1.95	2.25	3.375	5.0625	2.925
	2.0	1.6	3.2	4.0	8.0	16.0	6.4
	2.5	2.0	5.0	6.25	15.625	39.0625	12.5
	3.0	2.7	8.1	9.0	27.0	81.0	24.3
	3.5	3.4	11.9	12.25	42.875	150.0625	41.65
	4.0	4.1	16.4	16	64	256	65.6
Σ	17.5	16.2	47.65	50.75	161.875	548.1875	154.475

Here, $n = 7$, $\Sigma x = 17.5$, $\Sigma y = 16.2$, $\Sigma xy = 47.65$, $\Sigma x^2 = 50.75$, $\Sigma x^3 = 161.875$, $\Sigma x^4 = 548.1875$ and $\Sigma x^2 y = 154.475$

Substituting these values in the above normal equations, we get

$$16.2 = 50.75a + 17.5b + 7c$$
$$47.65 = 161.875a + 50.75b + 17.5c$$
$$154.475 = 548.1875a + 161.875b + 50.75c$$

Solving these equations, we get

$$a = 0.243,\quad b = -0.193 \text{ and } c = 1.04$$

$\therefore$ The equation of the best fitting curve is $y = 0.243x^2 - 0.193x + 1.04$

Example 4.5.9: Fit a second degree curve $y \, Z \, a \, H \, bx \, H \, cx^2$ for the data

x	**0**	**1**	**2**	**3**	**4**
y	N=	NKU=	**1.3**	**2.5**	**6.3**

Solution: **(VTU 2005, 2006, 2010)**

The normal equations of fitting $y = a + bx + cx^2$ are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	0	1	0	0	0	0	0
	1	1.8	1.8	1	1	1	1.8
	2	1.3	2.6	4	8	16	5.2
	3	2.5	7.5	9	27	81	22.5
	4	6.3	25.2	16	64	256	100.8
Σ	10	12.9	37.1	30	100	354	130.3

Here, $n = 5$, $\Sigma x = 10$, $\Sigma y = 12.9$, $\Sigma xy = 37.1$, $\Sigma x^2 = 30$, $\Sigma x^3 = 100$, $\Sigma x^4 = 354$ and $\Sigma x^2 y = 130.3$

Substituting these values in the above normal equations, we get

$$12.9 = 4a + 10b + 30c$$
$$37.1 = 10a + 30b + 100c$$
$$130.3 = 30a + 100b + 354c$$

Solving these equations, we get

$$a = 1.42,\quad b = -1.07 \text{ and } c = 0.55$$

$\therefore$ The equation of the best fitting curve is $y = 1.42 - 1.07x + 0.55\,x^2$

Example 4.5.10: Fit a second degree curve $y \, Z \, a \, H \, bx \, H \, cx^2$ for the data

x	J3	J2	J1	0	1	2	3
y	QKSP=	OKNN=	0.67	0.09	0.63	2.15	4.58

Solution: **(VTU 2006, 2009, 2012)**

The normal equations of fitting $y = a + bx + cx^2$ are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	–3	4.63	–13.89	9	–27	81	41.67
	–2	2.11	– 4.22	4	–8	16	8.44
	–1	0.67	– 0.67	1	–1	1	0.67
	0	0.09	0	0	0	0	0
	1	0.63	0.63	1	1	1	0.63
	2	2.15	4.3	4	8	16	8.6
	3	4.58	13.74	9	27	81	41.22
Σ	0	14.86	– 0.11	28	0	196	101.23

Here, $n = 7$, $\Sigma x = 0$, $\Sigma y = 14.86$, $\Sigma xy = -0.11$, $\Sigma x^2 = 28$, $\Sigma x^3 = 0$, $\Sigma x^4 = 196$ and $\Sigma x^2 y = 101.23$

Substituting these values in the above normal equations, we get

$$14.86 = 7a + 0b + 28c$$

$$-0.11 = 0a + 28b + 0c$$

$$101.23 = 28a + 0b + 196c$$

Solving these equations, we get

$$a = 0.1329, \quad b = -0.00393 \text{ and } c = 0.4975$$

$\therefore$ The equation of the best fitting curve is

$$y = 0.1329 - 0.00393x + 0.4975x^2$$

Example 4.5.11: Find the best values of *a*, *b* and *c* if the equation $y \; Z \; a \; H \; bx \; H \; cx^2$ is to fit most closely to the following observations:

x	J2	J1	0	1	2
y	JPKNRM=	J1.390	0.620	2.880	5.378

Solution: **(VTU 2007, 2013, 2016)**

The normal equations of fitting $y = a + bx + cx^2$ are

$$\Sigma y = na + b\Sigma x + c\Sigma x^2$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad \text{and}$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4$$

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	–2	-3.150	6.3	4	– 8	16	–12.6
	–1	–1.390	1.39	1	– 1	1	–1.39
	0	0.620	0	0	0	0	0
	1	2.880	2.880	1	1	1	2.880
	2	5.378	10.756	4	8	16	21.512
Σ	0	4.328	21.326	10	0	34	10.402

Here, n = 5, Σx = 0, Σy = 4.328, Σxy = 21.326, Σx^2 = 10, Σx^3 = 0, Σx^4 = 34 and $\Sigma x^2 y$ = 10.402

Substituting these values in the above normal equations, we get

$$4.338 = 5a + 0b + 10c$$
$$21.326 = 0a + 10b + 0c$$
$$10.402 = 10a + 0b + 34c$$

Solving these equations, we get

$$a = 0.621, \quad b = 2.1326 \text{ and } c = 0.1233$$

$\therefore$ The best values of a, b and c for fitting of $\boldsymbol{y \; Z \; a \; H \; bx \; H \; cx^2}$ are

$$a = 0.621, \quad b = 2.1326 \text{ and } c = 0.1233$$

Example 4.5.12: Fit a second degree curve $y \, Z \, ax^2 \, H \, bx \, H \, c$ for the data

x	1	2	3	4	5
y	NM=	12	13	16	19

Solution:

The normal equations of fitting $y = ax^2 + bx + c$ are

$$\therefore \quad \Sigma y = a\Sigma x^2 + b\Sigma x + nc$$
$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{and}$$
$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	1	10	10	1	1	1	10
	2	12	24	4	8	16	48
	3	13	39	9	27	81	117
	4	16	64	16	64	256	256
	5	19	95	25	125	625	475
Σ	15	61	232	55	225	979	906

Here, $n = 5$, $\Sigma x = 15$, $\Sigma y = 61$, $\Sigma xy = 232$, $\Sigma x^2 = 55$, $\Sigma x^3 = 225$, $\Sigma x^4 = 979$ and

$\Sigma x^2 y = 906$

Substituting these values in the above normal equations, we get

$$61 = 55a + 15b + 5c$$
$$232 = 225a + 55b + 15c$$
$$906 = 979a + 225b + 55c$$

Solving these equations, we get

$$a = 0.29, \quad b = 0.46 \text{ and } c = 9.43$$

$\therefore$ The equation of the best fitting curve is $y = 0.29x^2 + 0.46x + 9.43$

Example 4.5.13: Fit a second degree Parabola for the following data

x	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
y	0=	S=	**7**	**8**	**10**	**11**	**11**	**10**	**9**

Solution: **(VTU 2000, 2007, 2013)**

The normal equations of fitting a second degree parabola $y = ax^2 + bx + c$ are

$$\therefore \quad \Sigma y = a\Sigma x^2 + b\Sigma x + nc$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{and}$$

$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

We prepare a relevant table as follows

	x	y	xy	x^2	x^3	x^4	x^2y
	1	2	2	1	1	1	2
	2	6	12	4	8	16	24
	3	7	21	9	27	81	63
	4	8	32	16	64	256	128
	5	10	50	25	125	625	250
	6	11	66	36	216	1296	396
	7	11	77	49	343	2401	539
	8	10	80	64	512	4096	640
	9	9	81	81	729	6561	729
Σ	45	74	421	285	2025	15333	2771

Here, $n = 9$, $\Sigma x = 45$, $\Sigma y = 74$, $\Sigma xy = 421$, $\Sigma x^2 = 285$, $\Sigma x^3 = 2025$, $\Sigma x^4 = 15333$ and $\Sigma x^2 y = 2771$

Substituting these values in the above normal equations, we get

$$74 = 285a + 45b + 9c$$

$$421 = 2025a + 285b + 45c$$

$$2771 = 15333a + 2025b + 285c$$

Solving these equations, we get

$$a = -0.2673, \quad b = 3.523 \text{ and } c = -0.9282$$

$\therefore$ The equation of the best fitting curve is

$$y = -0.2673x^2 + 3.523x - 0.9282$$

EXERCISE 4.5

1. Fit a second degree Parabola to the following data:

x	1	2	3	4	5	6	7	8	9	10
y	124	129	140	159	228	289	315	302	263	210

2. Fit a second degree Parabola to the following data:

x	2	4	6	8	10
y	3.07	12.85	31.47	57.38	91.29

(VTU 2011)

3. Find the best values of a, b, c if the equation $y = a + bx + cx^2$ is to fit most closely to the following observation.

x	1	2	3	4	5
y	10	12	13	16	19

(VTU 2013)

4. The velocity V of a liquid is known to vary with temperature according to a quadratic law $V = a + bT + cT^2$. Find the best values of a, b and c for the data:

T	1	2	3	4	5	6	7
V	2.31	2.01	3.80	1.66	1.55	1.47	1.41

5. The following table gives the results of the measurements of train resistances, V is the velocity in miles per hour, R is the resistance in pounds per ton:

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46.0

If R is related to V by the relation $R = a + bV + cV^2$, find a, b and c.

6. Fit a second degree Parabola $y = ax^2 + bx + c$ to the following data:

x	10	20	30	40	50	60
y	157	179	210	252	302	361

ANSWERS

1. $y = 18.866 + 66.158x - 4.333x^2$
2. $y = 0.34 - 0.78x + 0.99x^2$
3. $y = 9.43 + 0.46x + 0.29x^2$
4. $a = 2.593, b = -0.326$ and $c = 0.023$
5. $a = 3.48, b = -0.002$ and $c = 0.0029$
6. $y = 0.046x^2 - 0.84x + 143.67$

❉❉❉❉❉❉