

## 7.3 FREQUENCY CHARTS

---

### 7.3.1 Variable

A quantity which can vary from one individual to another is called a **variable**. It is also called a **variate**. Wages, barometer readings, rainfall records, heights, and weights are the common examples of variables.

Quantities which can take any numerical value within a certain range are called **continuous variables**. For example, the height of a child at various ages is a continuous variable since, as the child grows from 120 cm to 150 cm, his height assumes all possible values within the limit.

Quantities which are incapable of taking all possible values are called **discontinuous** or **discrete variables**. For example, the number of rooms in a house can take only the integral values such as 2, 3, 4, etc.

### 7.3.2 Frequency Distributions

The scores of 50 students in mathematics are arranged below according to their roll numbers, the maximum scores being 100.

19, 70, 75, 15, 0, 23, 59, 56, 27, 89, 91, 22, 21, 22, 50, 89, 56, 73, 56, 89, 75, 65, 85, 22, 3, 12, 41, 87, 82, 72, 50, 22, 87, 50, 89, 28, 89, 50, 40, 36, 40, 30, 28, 87, 81, 90, 22, 15, 30, 35.

The data given in the crude form (or raw form) is called **ungrouped data**. If the data is arranged in ascending or descending order of magnitude, it is said to be arranged in an array. Let us now arrange it in the intervals 0–10, 10–20, 20–30, 30–40, 40–50, 50–60, 60–70, 70–80, 80–90, 90–100. This is arranged by a method called the **tally method**.

In this we consider every observation and put it in the suitable class by drawing a vertical line. After every 4 vertical lines, we cross it for the 5th entry and then a little space is left and the next vertical line is drawn.

<i>Scores</i> (Class-interval)	<i>Number of Students</i>	<i>Frequency</i> ( <i>f</i> )	<i>Cumulative</i> <i>Frequencies</i>
0—10		2	2
10—20		4	6
20—30		10	16
30—40		4	20
40—50		3	23
50—60		8	31
60—70		1	32
70—80		5	37
80—90		11	48
90—100		2	50
Total		$\Sigma f = 50$	

This type of representation is called a **grouped frequency distribution** or simply a **frequency distribution**. The groups are called the **classes** and the boundary ends 0, 10, 20, ..... etc. are called **class limits**. In the class limits 10—20, 10 is the **lower limit** and 20 is the **upper limit**. The difference between the upper and lower limits of a class is called its magnitude or **class-interval**. The number of observations falling within a particular class is called its **frequency** or **class frequency**. The frequency of the class 80—90 is 11. The variate value which lies mid-way between the upper and lower limits is called mid-value or mid-point of that class. The mid-points of these are respectively 5, 15, 25, 35, ..... The **cumulative frequency** corresponding to a class is the total of all the frequencies up to and including that class. Thus the cumulative frequency of the class 10—20 is 2 + 4, *i.e.*, 6 the cumulative frequency of the class 20—50 is 6 + 10, *i.e.*, 16, and so on.

While preparing the frequency distribution the following points must be remembered:

1. The class-intervals should be of equal width as far as possible. A comparison of different distributions is facilitated if the class interval is used for all. The class-interval should be an integer as far as possible.
2. The number of classes should never be fewer than 6 and not more than 30. With a smaller number of classes, the accuracy may be lost, and with a larger number of classes, the computations become tedious.
3. The observation corresponding to the common point of two classes should always be put in the higher class. For example, a number corresponding to the value 30 is to be put up in the class 30—40 and not in 20—30.

The following forms of the above table may also be used:

<i>Cumulative Frequency</i>			
<i>Scores</i>	<i>Number of Students</i>	<i>Scores</i>	<i>Number of Students</i>
Under 10	2	above 90	2
Under 20	6	above 80	13
Under 30	16	above 70	18
Under 40	20	above 60	19
Under 50	23	above 50	27
Under 60	31	above 40	30
Under 70	32	above 30	34
Under 80	37	above 20	44
Under 90	48	above 10	48
Under 100	50	above 0	50

## 7.4 GRAPHICAL REPRESENTATION OF A FREQUENCY DISTRIBUTION

Representation of frequency distribution by means of a diagram makes the unwieldy data intelligible and conveys to the eye the general run of the observations. The graphs and diagrams have a more lasting effect on the brain. It is always easier to compare data through graphs and diagrams. Forecasting also becomes easier with the help of graphs. Graphs help us in interpolation of values of the variables.

However there are certain disadvantages as well. Graphs do not give measurements of the variables as accurate as those given by tables. The numerical value can be obtained to any number of decimal places in a table, but from graphs it can not be found to 2nd or 3rd places of decimals. Another disadvantage is that it is very difficult to have a proper selection of scale. The facts may be misrepresented by differences in scale.

## 7.5 TYPES OF GRAPHS AND DIAGRAMS

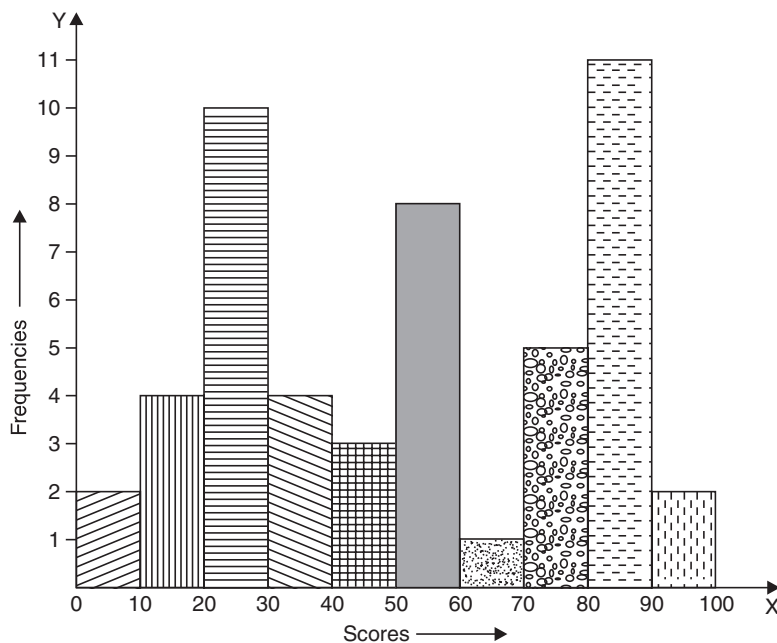
Generally the following types of graphs are used in representing frequency distributions:

(1) Histograms, (2) Frequency Polygon, (3) Frequency Curve, (4) Cumulative Frequency Curve or the Ogive, (5) Histograms, (6) Bar Diagrams, (7) Area

Diagrams, (8) Circles or Pie Diagrams, (9) Prisms, (10) Cartograms and Map Diagrams, (11) Pictograms.

## 7.6 HISTOGRAMS

To draw the histograms of a given grouped frequency distribution, mark off along a horizontal base line all the class-intervals on a suitable scale. With the class-intervals as bases, draw rectangles with the areas proportional to the frequencies of the respective class-intervals. For equal class-intervals, the heights of the rectangles will be proportional to the frequencies. If the class-intervals are not equal, the heights of the rectangles will be proportional to the ratios of the frequencies to the width of the corresponding classes. A diagram with all these rectangles is a **Histogram**.

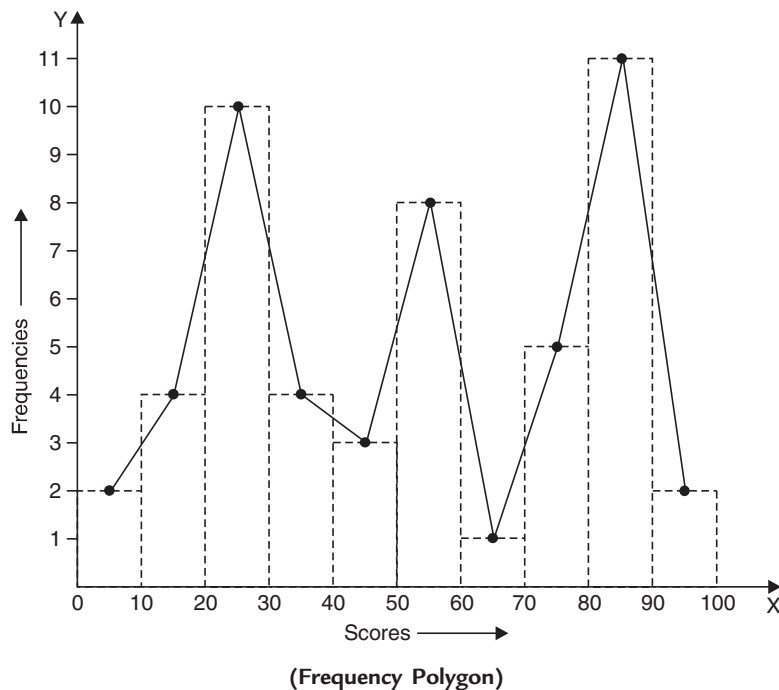


(Histogram for the previous table)

Histograms are also useful when the class-intervals are not of the same width. They are appropriate to cases in which the frequency changes rapidly.

## 7.7 FREQUENCY POLYGON

If the various points are obtained by plotting the central values of the class intervals as  $x$  co-ordinates and the respective frequencies as the  $y$  co-ordinates, and these points are joined by straight lines taken in order, they form a polygon called **Frequency Polygon**.



In a frequency polygon the variables or individuals of each class are assumed to be concentrated at the mid-point of the class-interval.

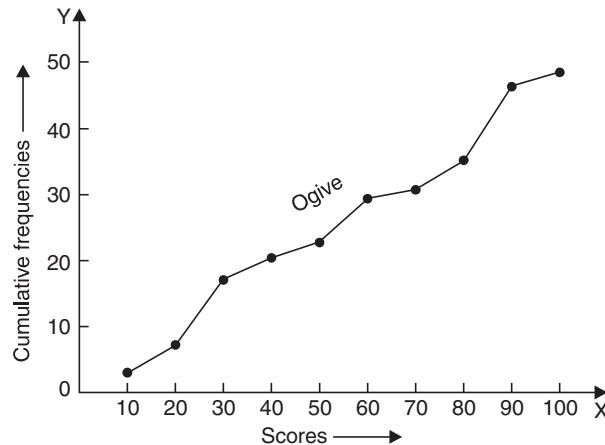
Here in this diagram dotted is the **Histogram** and a polygon with lines as sides is the **Frequency Polygon**.

## 7.8 FREQUENCY CURVE

If through the vertices of a frequency polygon a smooth freehand curve is drawn, we get the **Frequency Curve**. This is done usually when the class-intervals are of small widths.

## 7.9 CUMULATIVE FREQUENCY CURVE OR THE OGIVE

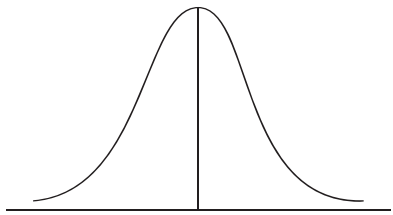
If from a cumulative frequency table, the upper limits of the class taken as  $x$  co-ordinates and the cumulative frequencies as the  $y$  co-ordinates and the points are plotted, then these points when joined by a freehand smooth curve give the **Cumulative Frequency Curve or the Ogive**.



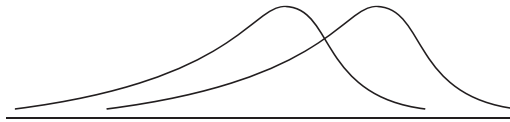
## 7.10 TYPES OF FREQUENCY CURVES

Following are some important types of frequency curves, generally obtained in the graphical representations of frequency distributions:

1. *Symmetrical curve or bell shaped curve.*
  2. *Moderately asymmetrical or skewed curve.*
  3. *Extremely asymmetrical or J-shaped curve or reverse J-shaped.*
  4. *U-shaped curve.*
  5. *A bimodal frequency curve.*
  6. *A multimodal frequency curve.*
1. **Symmetrical curve or Bell shaped curve.** If a curve can be folded symmetrically along a vertical line, it is called a symmetrical curve. In this type the class frequencies decrease to zero symmetrically on either side of a central maximum, *i.e.*, the observations equidistant from the central maximum have the same frequency.

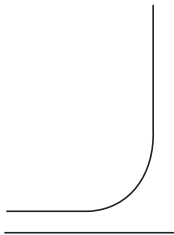


(Bell shaped curve)

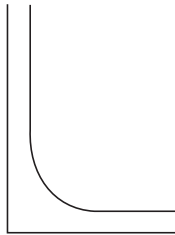


(Skewed curve)

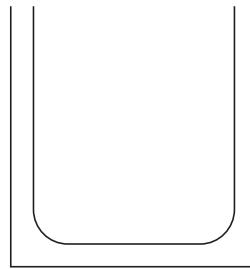
2. **Moderately asymmetrical or skewed curve.** If there is no symmetry in the curve, it is called a **Skew Curve**. In this case the class frequencies decrease with greater rapidity on one side of the maximum than on the other. In this curve one tail is always longer than the other. If the long tail is to the to be a positive side, it is said to be a positive skew curve, if long tail is to the negative side, it is said to be a negative skew curve.
3. **Extremely asymmetrical or J-shaped curve.** When the class frequencies run up to a maximum at one end of the range, they form a J-shaped curve.



J-shaped curve

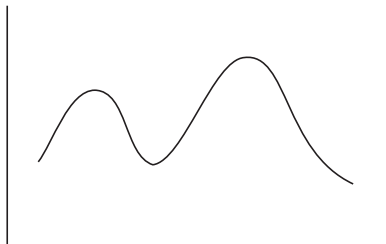


Reversed J-shaped curve

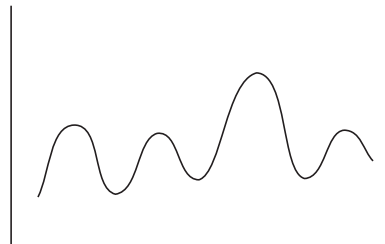


U-shaped curve

4. **U-shaped curve.** In this curve, the maximum frequency is at the ends of the range and a maximum towards the center.
5. A Bimodal curve has two maxima.



Bimodal curve

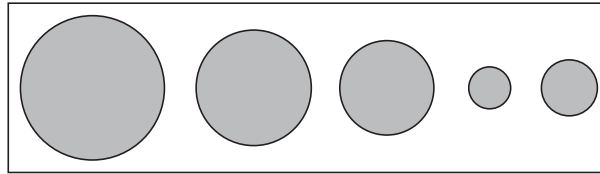


Multimodal curve

6. A multimodal curve has more than two maxima.

## 7.11 DIAGRAMS

1. **Bar diagrams.** Bar diagrams are used to compare the simple magnitude of different items. In bar diagrams, equal bases on a horizontal or vertical line are selected and rectangles are constructed with the length proportional to the given data. The width of bars is an arbitrary factor. The distance between two bars should be taken at about one-half of the width of a bar.
2. **Area diagrams.** When the difference between two quantities to be compared is large, bars do not show the comparison so clearly. In such cases, squares or circle are used.
3. **Circle or Pie-diagrams.** When circles are drawn to represent an area equivalent to the figures, they are said to form pie-diagrams or circles-diagrams. In case of circles, the square roots of magnitudes are proportional to the radii.



4. **Subdivided Pie-diagram.** Subdivided Pie-diagrams are used when comparison of the component parts is done with another and the total. The total value is equated to  $360^\circ$  and then the angles corresponding to the component parts are calculated.
5. **Prisms and Cubes.** When the ratio between the two quantities to be compared is very great so that even area diagrams are not suitable, the data can be represented by spheres, prisms, or cubes. Cubes are in common use. Cubes are constructed on sides which are taken in the ratio of cube roots of the given quantities.
6. **Cartograms or map diagrams.** Cartograms or map diagrams are most suitable for geographical data. Rainfalls and temperature in different parts of the country are shown with dots or shades in a particular map.
7. **Pictograms.** When numerical data are represented by pictures, they give a more attractive representation. Such pictures are called pictograms.



## 7.12 CURVE FITTING

---

Let there be two variables  $x$  and  $y$  which give us a set of  $n$  pairs of numerical values  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . In order to have an approximate idea about the relationship of these two variables, we plot these  $n$  paired points on a graph, thus we get a diagram showing the simultaneous variation in values of both the variables called *scatter or dot diagram*. From scatter diagram, we get only an approximate non-mathematical relation between two variables. *Curve fitting* means an exact relationship between two variables by algebraic equations. In fact, this relationship is the equation of the curve. Therefore, *curve fitting* means to form an equation of the curve from the given data. Curve fitting is considered of immense importance both from the point of view of theoretical and practical statistics.

Theoretically, curve fitting is useful in the study of correlation and regression. Practically, it enables us to represent the relationship between two variables by simple algebraic expressions, for example, polynomials, exponential, or logarithmic functions.

Curve fitting is also used to estimate the values of one variable corresponding to the specified values of the other variable.

The constants occurring in the equation of an approximate curve can be found by the following methods:

- (i) Graphical method
- (ii) Method of group averages
- (iii) Principle of least squares
- (iv) Method of moments.

Out of the above four methods, we will only discuss and study here *the principle of least squares*.

## 7.13 PRINCIPLE OF LEAST SQUARES

---

Principle of least squares provides a unique set of values to the constants and hence suggests a curve of best fit to the given data.

Suppose we have  $m$ -paired observations  $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$  of two variables  $x$  and  $y$ . It is required to fit a polynomial of degree  $n$  of the type

$$y = a + bx + cx^2 + \dots + kx^n \quad (1)$$

of these values. We have to determine the constants  $a, b, c, \dots, k$  such that they represent the curve of best fit of that degree.

In case  $m = n$ , we get in general a unique set of values satisfying the given system of equations.

But if  $m > n$ , then we get  $m$  equations by putting different values of  $x$  and  $y$  in equation (1) and we want to find only the values of  $n$  constants. Thus there may be no such solution to satisfy all  $m$  equations.

Therefore we try to find out those values of  $a, b, c, \dots, k$  which satisfy all the equations as nearly as possible. We apply the principle of least squares in such cases.

Putting  $x_1, x_2, \dots, x_m$  for  $x$  in (1), we get

$$\begin{aligned} y_1' &= a + bx_1 + cx_1^2 + \dots + kx_1^n \\ y_2' &= a + bx_2 + cx_2^2 + \dots + kx_2^n \\ &\vdots \\ y_m' &= a + bx_m + cx_m^2 + \dots + kx_m^n \end{aligned}$$

where  $y_1', y_2', \dots, y_m'$  are the expected values of  $y$  for  $x = x_1, x_2, \dots, x_m$  respectively.

The values  $y_1, y_2, \dots, y_m$  are called observed values of  $y$  corresponding to  $x = x_1, x_2, \dots, x_m$  respectively.

The expected values are different from the observed values, the difference  $y_r - y_r'$  for different values of  $r$  are called *residuals*.

Introduce a new quantity  $U$  such that

$$U = \Sigma(y_r - y_r')^2 = \Sigma(y_r - a - bx_r - cx_r^2 - \dots - kx_r^n)^2$$

The constants  $a, b, c, \dots, k$  are chosen in such a way that the sum of the squares of the residuals is minimum.

Now the condition for  $U$  to be maximum or minimum is  $\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} = \dots = \frac{\partial U}{\partial k}$ . On simplifying these relations, we get

$$\begin{aligned} \Sigma y &= ma + b\Sigma x + \dots + k\Sigma x^n \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + \dots + k\Sigma x^{n+1} \\ \Sigma x^2y &= a\Sigma x^2 + b\Sigma x^3 + \dots + k\Sigma x^{n+2} \\ &\vdots \\ \Sigma x^ny &= a\Sigma x^n + b\Sigma x^{n+1} + \dots + k\Sigma x^{2n} \end{aligned}$$

These are known as *Normal equations* and can be solved as simultaneous equations to give the values of the constants  $a, b, c, \dots, k$ . These equations are  $(n + 1)$  in number.

If we calculate the second order partial derivatives and these values are given, they give a positive value of the function, so  $U$  is minimum.

This method does not help us to choose the degree of the curve to be fitted but helps us in finding the values of the constants when the form of the curve has already been chosen.

### 7.14 FITTING A STRAIGHT LINE

Let  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$  be  $n$  sets of observations of related data and

$$y = a + bx \quad (2)$$

be the straight line to be fitted. The residual at  $x = x_i$  is

$$E_i = y_i - f(x_i) = y_i - a - bx_i$$

Introduce a new quantity  $U$  such that

$$U = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

By the principle of Least squares,  $U$  is minimum

$$\therefore \frac{\partial U}{\partial a} = 0 \quad \text{and} \quad \frac{\partial U}{\partial b} = 0$$

$$\therefore 2 \sum_{i=1}^n (y_i - a - bx_i)(-1) = 0 \quad \text{or} \quad \boxed{\Sigma y = na + b \Sigma x} \quad (3)$$

$$\text{and} \quad 2 \sum_{i=1}^n (y_i - a - bx_i)(-x_i) = 0 \quad \text{or} \quad \boxed{\Sigma xy = a \Sigma x + b \Sigma x^2} \quad (4)$$

Since  $x_i, y_i$  are known, equations (3) and (4) result in  $a$  and  $b$ . Solving these, the best values for  $a$  and  $b$  can be known, and hence equation (2).

NOTE

*In case of change of origin,*

*if  $n$  is odd then,*

$$u = \frac{x - (\text{middle term})}{\text{interval } (h)}$$

*but if  $n$  is even then*

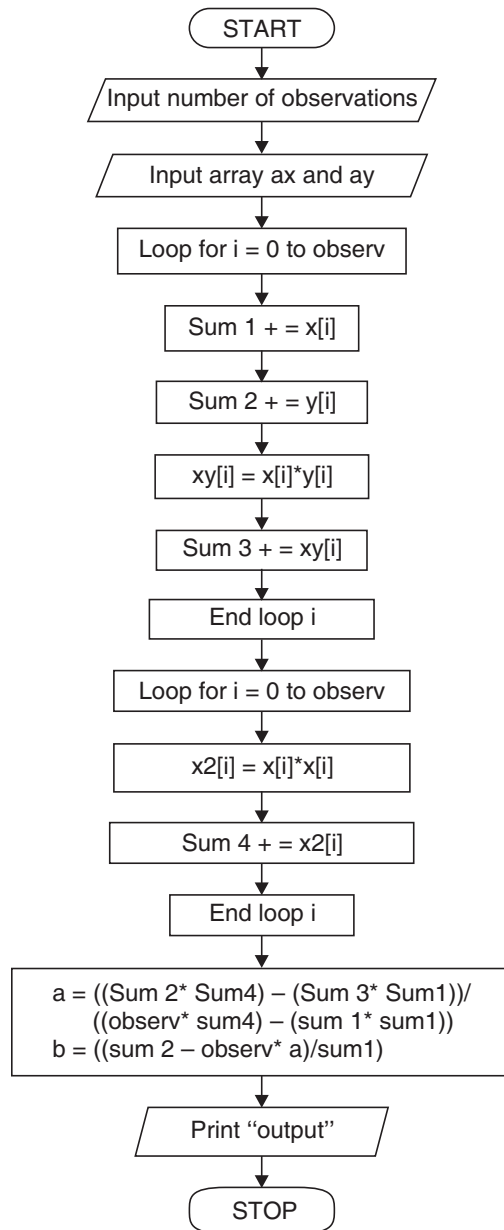
$$u = \frac{x - (\text{mean of two middle terms})}{\frac{1}{2}(\text{interval})}$$

### 7.15 ALGORITHM FOR FITTING A STRAIGHT LINE OF THE FORM $y = a + bx$ FOR A GIVEN SET OF DATA POINTS

---

- Step 01.** Start of the program.
- Step 02.** Input no. of terms observ
- Step 03.** Input the array ax
- Step 04.** Input the array ay
- Step 05.** for i=0 to observ
- Step 06.** sum1+=x[i]
- Step 07.** sum2+=y[i]
- Step 08.** xy[i]=x[i]\*y[i];
- Step 09.** sum3+=xy[i]
- Step 10.** End Loop i
- Step 11.** for i = 0 to observ
- Step 12.** x2[i]=x[i]\*x[i]
- Step 13.** sum4+=x2[i]
- Step 14.** End of Loop i
- Step 15.** temp1=(sum2\*sum4)-(sum3\*sum1)
- Step 16.** a=temp1/((observ \*sum4)-(sum1\*sum1))
- Step 17.** b=(sum2-observ\*a)/sum1
- Step 18.** Print output a,b
- Step 19.** Print "line is:  $y = a+bx$ "
- Step 20.** End of Program

### 7.16 FLOW-CHART FOR FITTING A STRAIGHT LINE $y = a + bx$ FOR A GIVEN SET OF DATA POINTS



```
/* *****
```

## 7.17 PROGRAM TO IMPLEMENT CURVE FITTING TO FIT A STRAIGHT LINE

---

```
***** */
```

```
//... HEADER FILE DECLARATION
# include <stdio.h>
# include <conio.h>
# include <math.h>
//... Main Execution Thread
void main()
{
//... Variable Declaration Field
//... Integer Type
int i=0;
int observ;
//... Floating Type
float x[10];
float y[10];
float xy[10];
float x2[10];
float sum1=0.0;
float sum2=0.0;
float sum3=0.0;
float sum4=0.0;
//... Double Type
double a;
double b;
//... Invoke Function Clear Screen
clrscr ();
//... Input Section
//... Input Number of Observations
printf("\n\n Enter the number of observations - ");
scanf("%d" ,&observ);
```

```

//... Input Sequel For Array X
printf("\n\n\n Enter the values of x - \n");
for (;i<observ;i++)
{
printf("\n\n Enter the Value of x%d: ",i+1);
scanf("%f" ,&x[i]);
sum1 +=x[i];
}
//... Input Sequel For Array Y
printf("\n\n Enter the values of y - \n");
for(i=0;i<observ;i++)
{
printf("\n\n Enter the value of y%d:",i+1);
scanf("%f",&y[i]);
sum2+=y[i];
}
//... Processing and Calculation Section
for(i=0;i<observ;i++)
{
xy[i]=x[i]*y[i];
sum3 +=xy[i];
}
for(i=0;i<observ; i++)
{
x2[i]=x[i]*x[i];
sum4+ =x2[i];
}
a=(sum2*sum4-sum3*sum1)/(observ*sum4-sum1*sum1);
b=(sum2-observ*a)/sum1;
//... Output Section
printf("\n\n\n\n Equation of the STRAIGHT LINE");
printf("of the form y = a + b*x is:");
printf("\n\n\n \t\t\t Y = %.2f + (%.2f) X", a,b);
//... Invoke User Watch Halt Function

```

```
printf("\n\n\n Press Enter to Exit");
getch();
}
//... Termination of Main Execution Thread
```

### EXAMPLES

**Example 1.** By the method of least squares, find the straight line that best fits the following data:

$x:$	1	2	3	4	5
$y:$	14	27	40	55	68.

**Sol.** Let the straight line of best fit be

$$y = a + bx \quad (5)$$

Normal equations are  $\Sigma y = ma + b\Sigma x$  (6)

and  $\Sigma xy = a\Sigma x + b\Sigma x^2$  (7)

Here  $m = 5$

The table is as below:

$x$	$y$	$xy$	$x^2$
1	14	14	1
2	27	54	4
3	40	120	9
4	55	220	16
5	68	340	25
$\Sigma x = 15$	$\Sigma y = 204$	$\Sigma xy = 748$	$\Sigma x^2 = 55$

Substituting in (6) and (7), we get

$$204 = 5a + 15b$$

$$748 = 15a + 55b$$

Solving, we get  $a = 0$ ,  $b = 13.6$

Hence required straight line is  $y = 13.6x$

**Example 2.** Fit a straight line to the following data:

$x:$	0	1	2	3	4
$y:$	1	1.8	3.3	4.5	6.3.



**Sol.** Let the straight line obtained from the given data be  $y = a + bx$  then the normal equations are

$$\Sigma y = ma + b \Sigma x \quad (8)$$

$$\Sigma xy = a \Sigma x + b \Sigma x^2 \quad (9)$$

Here  $m = 5$

$x$	$y$	$xy$	$x^2$
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x = 10$	$\Sigma y = 16.9$	$\Sigma xy = 47.1$	$\Sigma x^2 = 30$

From (8) and (9),  $16.9 = 5a + 10b$

and  $47.1 = 10a + 30b$

Solving, we get  $a = 0.72, b = 1.33$

$\therefore$  Required line is  $y = 0.72 + 1.33x$ .

**Example 3.** Fit a straight line to the following data regarding  $x$  as the independent variable:

$x$ :	1	2	3	4	5	6
$y$ :	1200	900	600	200	110	50.

**Sol.** Let the equation of the straight line to be fitted be  $y = a + bx$

Here  $m = 6$

$x$	$y$	$x^2$	$xy$
1	1200	1	1200
2	900	4	1800
3	600	9	1800
4	200	16	800
5	110	25	550
6	50	36	300
$\Sigma x = 21$	$\Sigma y = 3060$	$\Sigma x^2 = 91$	$\Sigma xy = 6450$

From normal equations, we get

$$3060 = 6a + 21b, 6450 = 21a + 91b$$

Solving, we get  $a = 1361.97, b = -243.42$

$\therefore$  Required line is

$$y = 1361.97 - 243.42x.$$

**Example 4.** Show that the line of fit to the following data is given by  $y = 0.7x + 11.285$ :

$x$ :	0	5	10	15	20	25
$y$ :	12	15	17	22	24	30.

**Sol.** Since  $m$  is even,

Let  $x_0 = 12.5$   $h = 5$   $y_0 = 20$  (say)

Then let,  $u = \frac{x - 12.5}{2.5}$  and  $v = y - 20$

$x$	$y$	$u$	$v$	$uv$	$u^2$
0	12	-5	-8	40	25
5	15	-3	-5	15	9
10	17	-1	-3	3	1
15	22	1	2	2	1
20	24	3	4	12	9
25	30	5	10	50	25
Total		$\Sigma u = 0$	$\Sigma v = 0$	$\Sigma uv = 122$	$\Sigma u^2 = 70$

Normal equations are  $0 = 6a$  and  $122 = 70b$

$\Rightarrow a = 0, b = 1.743$

Line of fit is  $v = 1.743u$

Put  $u = \frac{x - 12.5}{2.5}$  and  $v = y - 20$ , we get

$$y = 0.7x + 11.285.$$

**Example 5.** Fit a straight line to the following data:

$x$ :	71	68	73	69	67	65	66	67
$y$ :	69	72	70	70	68	67	68	64.

**Sol.** Let the equation of the straight line to be fitted be

$$y = a + bx \quad (10)$$

Normal equations are

$$\Sigma y = ma + b\Sigma x \quad (11)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad (12)$$

Here  $m = 8$ . Table is as below:

$x$	$y$	$xy$	$x^2$
71	69	4899	5041
68	72	4896	4624
73	70	5110	5329
69	70	4830	4761
67	68	4556	4489
65	67	4355	4225
66	68	4488	4356
67	64	4288	4489
$\Sigma x = 546$	$\Sigma y = 548$	$\Sigma xy = 37422$	$\Sigma x^2 = 37314$

Substituting these values in equations (11) and (12), we get

$$548 = 8a + 546b$$

$$37422 = 546a + 37314b$$

Solving, we get

$$a = 39.5454, \quad b = 0.4242$$

Hence the required line of best fit is

$y = 39.5454 + 0.4242 x.$

**Example 6.** Show that the best fitting linear function for the points  $(x_1, y_1)$ ,  $(x_2, y_2)$ , ...,  $(x_n, y_n)$  may be expressed in the form

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0 \quad (i = 1, 2, \dots, n)$$

Show that the line passes through the mean point  $(\bar{x}, \bar{y})$ .

**Sol.** Let the best fitting linear function be  $y = a + bx$  (13)

Then the normal equations are

$$\Sigma y_i = na + b\Sigma x_i \quad (14)$$

and

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \quad (15)$$

Equations (13), (14), (15) may be rewritten as

$$bx - y + a = 0$$

$$b\Sigma x_i - \Sigma y_i + na = 0$$

and

$$b\Sigma x_i^2 - \Sigma x_i y_i + a\Sigma x_i = 0$$

Eliminating  $a$  and  $b$  between these equations

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0 \quad (16)$$

which is the required best fitting linear function for the mean point  $(\bar{x}, \bar{y})$ ,

$$\bar{x} = \frac{1}{n} \Sigma x_i \quad \bar{y} = \frac{1}{n} \Sigma y_i.$$

Clearly, the line (16) passes through point  $(\bar{x}, \bar{y})$  as two rows of determinants being equal make it zero.

### ASSIGNMENT 7.1

1. Fit a straight line to the given data regarding  $x$  as the independent variable:

$x$	1	2	3	4	6	8
$y$	2.4	3.1	3.5	4.2	5.0	6.0

2. Find the best values of  $a$  and  $b$  so that  $y = a + bx$  fits the given data:

$x$	0	1	2	3	4
$y$	1.0	2.9	4.8	6.7	8.6

3. Fit a straight line approximate to the data:

$x$	1	2	3	4
$y$	3	7	13	21

4. A simply supported beam carries a concentrated load  $P(lb)$  at its mid-point. Corresponding to various values of  $P$ , the maximum deflection  $Y$  (*in*) is measured. The data are given below. Find a law of the type  $Y = a + bP$

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

5. In the following table  $y$  is the weight of potassium bromide which will dissolve in 100 grams of water at temperature  $x^\circ$ . Find a linear law between  $x$  and  $y$

$x^\circ(c)$	0	10	20	30	40	50	60	70
$y \text{ gm}$	53.5	59.5	65.2	70.6	75.5	80.2	85.5	90

6. The weight of a calf taken at weekly intervals is given below. Fit a straight line using the method of least squares and calculate the average rate of growth per week.

Age	1	2	3	4	5	6	7	8	9	10
Weight	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

7. Find the least square line for the data points  
 $(-1, 10)$ ,  $(0, 9)$ ,  $(1, 7)$ ,  $(2, 5)$ ,  $(3, 4)$ ,  $(4, 3)$ ,  $(5, 0)$  and  $(6, -1)$ .
8. Find the least square line  $y = a + bx$  for the data:

$x_i$	-2	-1	0	1	2
$y_i$	1	2	3	3	4

9. 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 data:

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

where  $P$  and  $W$  are taken in kg-wt.

10. Using the method of least squares, fit a straight line to the following data:

$x$	1	2	3	4	5
$y$	2	4	6	8	10

11. Differentiate between interpolating polynomial and least squares polynomial obtained for a set of data.

### 7.18 FITTING OF AN EXPONENTIAL CURVE $y = ae^{bx}$

---

Taking logarithms on both sides, we get

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

i.e.,

$$\boxed{Y = A + Bx} \quad (17)$$

where  $Y = \log_{10} y$ ,  $A = \log_{10} a$  and  $B = b \log_{10} e$

The normal equations for (17) are  $\Sigma Y = nA + B\Sigma x$  and  $\Sigma xY = A\Sigma x + B\Sigma x^2$

Solving these, we get A and B.

Then  $a = \text{antilog } A$  and  $b = \frac{B}{\log_{10} e}$ .

### 7.19 FITTING OF THE CURVE $y = ax^b$

---

Taking the logarithm on both sides, we get

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

i.e.,

$$\boxed{Y = A + bX} \quad (18)$$

where  $Y = \log_{10} y$ ,  $A = \log_{10} a$  and  $X = \log_{10} x$ .

The normal equations to (18) are  $\Sigma Y = nA + b\Sigma X$

and  $\Sigma XY = A\Sigma X + b\Sigma X^2$

which results A and b on solving and  $a = \text{antilog } A$ .

### 7.20 FITTING OF THE CURVE $y = ab^x$

---

Take the logarithm on both sides,

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

$$\Rightarrow Y = A + Bx$$

where  $Y = \log y$ ,  $A = \log a$ ,  $B = \log b$ .

This is a linear equation in Y and x.

For estimating A and B, normal equations are

$$\Sigma Y = nA + B \Sigma x$$

and

$$\Sigma xY = A \Sigma x + B \Sigma x^2$$

where  $n$  is the number of pairs of values of  $x$  and  $y$ .

Ultimately,  $a = \text{antilog}(A)$  and  $b = \text{antilog}(B)$ .

## 7.21 FITTING OF THE CURVE $pv^r = k$

---

$$pv^r = k \Rightarrow v = k^{1/r} p^{-1/r}$$

Taking logarithm on both sides,

$$\log v = \frac{1}{r} \log k - \frac{1}{r} \log p$$

$\Rightarrow$

$$Y = A + BX$$

where  $Y = \log v$ ,  $A = \frac{1}{r} \log k$ ,  $B = -\frac{1}{r}$  and  $X = \log p$

$r$  and  $k$  are determined by the above equations. Normal equations are obtained as per that of the straight line.

## 7.22 FITTING OF THE CURVE OF TYPE $xy = b + ax$

---

$$xy = b + ax \Rightarrow y = \frac{b}{x} + a$$

$\Rightarrow Y = bX + a$ , where  $X = \frac{1}{x}$ .

Normal equations are  $\Sigma Y = na + b \Sigma X$   
 $\Sigma XY = a \Sigma X + b \Sigma X^2$ .

## 7.23 FITTING OF THE CURVE $y = ax^2 + \frac{b}{x}$

---

Let the  $n$  points be  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

Error of estimate for  $i^{\text{th}}$  point  $(x_i, y_i)$  is

$$E_i = \left( y_i - ax_i^2 - \frac{b}{x_i} \right)$$

By principle of Least squares, the values of  $a$  and  $b$  are such that

$$U = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n \left( y_i - ax_i^2 - \frac{b}{x_i} \right)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow \sum_{i=1}^n x_i^2 y_i = a \sum_{i=1}^n x_i^4 + b \sum_{i=1}^n x_i$$

and  $\frac{\partial U}{\partial b} = 0$

$$\Rightarrow \sum_{i=1}^n \frac{y_i}{x_i} = a \sum_{i=1}^n x_i + b \sum_{i=1}^n \frac{1}{x_i^2}$$

or Dropping the suffix  $i$ , normal equations are

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

and  $\Sigma \frac{y}{x} = a \Sigma x + b \Sigma \frac{1}{x^2}.$

## 7.24 FITTING OF THE CURVE $y = ax + bx^2$

Error of estimate for  $i^{\text{th}}$  point  $(x_i, y_i)$  is  $E_i = (y_i - ax_i - bx_i^2)$

By the principle of Least Squares, the values of  $a$  and  $b$  are such that

$$U = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - ax_i - bx_i^2)^2 \text{ is minimum.}$$

Normal equations are given by  $\frac{\partial U}{\partial a} = 0$

$$\Rightarrow \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i^3$$

and  $\frac{\partial U}{\partial b} = 0$



$$\Rightarrow \sum_{i=1}^n x_i^2 y_i = a \sum_{i=1}^n x_i^3 + b \sum_{i=1}^n x_i^4$$

or Dropping the suffix  $i$ , normal equations are

$$\Sigma xy = a \Sigma x^2 + b \Sigma x^3$$

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

## 7.25 FITTING OF THE CURVE $y = ax + \frac{b}{x}$

Error of estimate for  $i^{\text{th}}$  point  $(x_i, y_i)$  is

$$E_i = y_i - ax_i - \frac{b}{x_i}$$

By the principle of Least Squares the values of  $a$  and  $b$  are such that

$$U = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n \left( y_i - ax_i - \frac{b}{x_i} \right)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow 2 \sum_{i=1}^n \left( y_i - ax_i - \frac{b}{x_i} \right) (-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + nb$$

and

$$\frac{\partial U}{\partial b} = 0$$

$$\Rightarrow 2 \sum_{i=1}^n \left( y_i - ax_i - \frac{b}{x_i} \right) \left( -\frac{1}{x_i} \right) = 0$$

$$\Rightarrow \sum_{i=1}^n \frac{y_i}{x_i} = na + b \sum_{i=1}^n \frac{1}{x_i^2}$$

Dropping the suffix  $i$ , normal equations are

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

and 
$$\Sigma \frac{y}{x} = na + b \Sigma \frac{1}{x^2}$$

where  $n$  is the number of pairs of values of  $x$  and  $y$ .

## 7.26 FITTING OF THE CURVE $y = a + \frac{b}{x} + \frac{c}{x^2}$

Normal equations are

$$\Sigma y = ma + b \Sigma \frac{1}{x} + c \Sigma \frac{1}{x^2}$$

$$\Sigma \frac{y}{x} = a \Sigma \frac{1}{x} + b \Sigma \frac{1}{x^2} + c \Sigma \frac{1}{x^3}$$

$$\Sigma \frac{y}{x^2} = a \Sigma \frac{1}{x^2} + b \Sigma \frac{1}{x^3} + c \Sigma \frac{1}{x^4}$$

where  $m$  is the number of pairs of values of  $x$  and  $y$ .

## 7.27 FITTING OF THE CURVE $y = \frac{c_0}{x} + c_1 \sqrt{x}$

Error of estimate for  $i^{\text{th}}$  point  $(x_i, y_i)$  is

$$E_i = y_i - \frac{c_0}{x_i} - c_1 \sqrt{x_i}$$

By the principle of Least Squares, the values of  $a$  and  $b$  are such that

$$U = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - \frac{c_0}{x_i} - c_1 \sqrt{x_i})^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial c_0} = 0 \quad \text{and} \quad \frac{\partial U}{\partial c_1} = 0$$

Now, 
$$\frac{\partial U}{\partial c_0} = 0$$

$$\Rightarrow 2 \sum_{i=1}^n \left( y_i - \frac{c_0}{x_i} - c_1 \sqrt{x_i} \right) \left( -\frac{1}{x_i} \right) = 0$$

$$\Rightarrow \sum_{i=1}^n \frac{y_i}{x_i} = c_0 \sum_{i=1}^n \frac{1}{x_i^2} + c_1 \sum_{i=1}^n \frac{1}{\sqrt{x_i}} \quad (19)$$

Also,  $\frac{\partial U}{\partial c_1} = 0$

$$\Rightarrow 2 \sum_{i=1}^n \left( y_i - \frac{c_0}{x_i} - c_1 \sqrt{x_i} \right) (-\sqrt{x_i}) = 0$$

$$\Rightarrow \sum_{i=1}^n y_i \sqrt{x_i} = c_0 \sum_{i=1}^n \frac{1}{\sqrt{x_i}} + c_1 \sum_{i=1}^n x_i \quad (20)$$

Dropping the suffix  $i$ , normal equations (19) and (20) become

$$\sum \frac{y}{x} = c_0 \sum \frac{1}{x^2} + c_1 \sum \frac{1}{\sqrt{x}}$$

and  $\sum y \sqrt{x} = c_0 \sum \frac{1}{\sqrt{x}} + c_1 \sum x.$

## 7.28 FITTING OF THE CURVE $2^x = ax^2 + bx + c$

Normal equations are

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

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

and  $\Sigma 2^x = a \Sigma x^2 + b \Sigma x + mc$

where  $m$  is number of points  $(x_i, y_i)$

### EXAMPLES

**Example 1.** Find the curve of best fit of the type  $y = ae^{bx}$  to the following data by the method of Least Squares:

$x:$	1	5	7	9	12
$y:$	10	15	12	15	21.

**Sol.** The curve to be fitted is  $y = ae^{bx}$

or  $Y = A + Bx$ ,

where  $Y = \log_{10} y$ ,  $A = \log_{10} a$ , and  $B = b \log_{10} e$

$\therefore$  The normal equations are  $\Sigma Y = 5A + B\Sigma x$

and  $\Sigma xY = A\Sigma x + B\Sigma x^2$

$x$	$y$	$Y = \log_{10} y$	$x^2$	$xY$
1	10	1.0000	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
$\Sigma x = 34$		$\Sigma Y = 5.7536$	$\Sigma x^2 = 300$	$\Sigma xY = 40.8862$

Substituting the values of  $\Sigma x$ , etc. calculated by means of above table in the normal equations.

We get  $5.7536 = 5A + 34B$

and  $40.8862 = 34A + 300B$

On solving  $A = 0.9766$ ;  $B = 0.02561$

$\therefore a = \text{antilog}_{10} A = 9.4754$ ;  $b = \frac{B}{\log_{10} e} = 0.059$

Hence the required curve is

$$y = 9.4754e^{0.059x}$$

**Example 2.** For the data given below, find the equation to the best fitting exponential curve of the form  $y = ae^{bx}$

$x$ :	1	2	3	4	5	6
$y$ :	1.6	4.5	13.8	40.2	125	300.

**Sol.**  $y = ae^{bx}$

Take log,  $\log y = \log a + bx \log e$

which is of the form  $Y = A + Bx$

where  $Y = \log y$ ,  $A = \log a$ ,  $B = b \log e$

$x$	$y$	$Y = \log y$	$x^2$	$xY$
1	1.6	.2041	1	.2041
2	4.5	.6532	4	1.3064
3	13.8	1.1399	9	3.4197
4	40.2	1.6042	16	6.4168
5	125	2.0969	25	10.4845
6	300	2.4771	36	14.8626
$\Sigma x = 21$		$\Sigma Y = 8.1754$	$\Sigma x^2 = 91$	$\Sigma xY = 36.6941$

Normal equations are

$$\text{and} \quad \left. \begin{aligned} \Sigma Y &= mA + B\Sigma x \\ \Sigma xY &= A\Sigma x + B\Sigma x^2 \end{aligned} \right\} \quad (21)$$

Here  $m = 6$

$$\therefore \text{From (21), } 8.1754 = 6A + 21B, \quad 36.6941 = 21A + 91B$$

$$\Rightarrow \quad A = -0.2534, \quad B = 0.4617$$

$$\therefore \quad \begin{aligned} a &= \text{antilog } A = \text{antilog } (-.2534) \\ &= \text{antilog } (\bar{1}.7466) = 0.5580 \end{aligned}$$

$$\text{and} \quad b = \frac{B}{\log e} = \frac{.4617}{.4343} = 1.0631$$

Hence required equation is

$$y = 0.5580 e^{1.0631x}$$

✓ **Example 3.** Determine the constants  $a$  and  $b$  by the Method of Least Squares such that  $y = ae^{bx}$  fits the following data:

$x$	2	4	6	8	10
$y$	4.077	11.084	30.128	81.897	222.62

**Sol.**  $y = ae^{bx}$

Taking log on both sides

$$\log y = \log a + bx \log e$$

or  $Y = A + BX,$

where  $Y = \log y$

$$A = \log a$$

$$B = b \log_{10} e$$

$$X = x.$$

Normal equations are

$$\Sigma Y = mA + B\Sigma X \quad (22)$$

and

$$\Sigma XY = A\Sigma X + B\Sigma X^2. \quad (23)$$

Here  $m = 5$ .

Table is as follows:

$x$	$y$	$X$	$Y$	$XY$	$X^2$
2	4.077	2	.61034	1.22068	4
4	11.084	4	1.04469	4.17876	16
6	30.128	6	1.47897	8.87382	36
8	81.897	8	1.91326	15.30608	64
10	222.62	10	2.347564	23.47564	100
		$\Sigma X = 30$	$\Sigma Y = 7.394824$	$\Sigma XY = 53.05498$	$\Sigma X^2 = 220$

Substituting these values in equations (22) and (23), we get

$$7.394824 = 5A + 30B$$

and

$$53.05498 = 30A + 220B.$$

Solving, we get

$$A = 0.1760594$$

and

$$B = 0.2171509$$

$\therefore$

$$a = \text{antilog}(A)$$


$$= \text{antilog}(0.1760594) = 1.49989$$

and

$$b = \frac{B}{\log_{10} e} = \frac{0.2171509}{.4342945} = 0.50001$$

Hence the required equation is

$$y = 1.49989 e^{0.50001x}.$$

 **Example 4.** Obtain a relation of the form  $y = ab^x$  for the following data by the Method of Least Squares:

$x$	2	3	4	5	6
$y$	8.3	15.4	33.1	65.2	126.4

**Sol.** The curve to be fitted is  $y = ab^x$

or  $Y = A + Bx$ ,

where  $A = \log_{10} a$ ,  $B = \log_{10} b$  and  $Y = \log_{10} y$ .

$\therefore$  The normal equations are  $\Sigma Y = 5A + B\Sigma x$

and  $\Sigma XY = A\Sigma x + B\Sigma x^2$ .

$x$	$y$	$Y = \log_{10} y$	$x^2$	$xY$
2	8.3	0.9191	4	1.8382
3	15.4	1.1872	9	3.5616
4	33.1	1.5198	16	6.0792
5	65.2	1.8142	25	9.0710
6	127.4	2.1052	36	12.6312
$\Sigma x = 20$		$\Sigma Y = 7.5455$	$\Sigma x^2 = 90$	$\Sigma xY = 33.1812$

Substituting the values of  $\Sigma x$ , etc. from the above table in normal equations, we get

$$7.5455 = 5A + 20B \quad \text{and} \quad 33.1812 = 20A + 90B.$$


On solving  $A = 0.31$  and  $B = 0.3$

$\therefore a = \text{antilog } A = 2.04$

and  $b = \text{antilog } B = 1.995$ .

Hence the required curve is

$$y = 2.04(1.995)^x.$$

 **Example 5.** By the method of least squares, find the curve  $y = ax + bx^2$  that best fits the following data:

$x$	1	2	3	4	5
$y$	1.8	5.1	8.9	14.1	19.8

**Sol.** Error of estimate for  $i^{\text{th}}$  point  $(x_i, y_i)$  is  $E_i = (y_i - ax_i - bx_i^2)$

By the principle of least squares, the values of  $a$  and  $b$  are such that

$$U = \sum_{i=1}^5 E_i^2 = \sum_{i=1}^5 (y_i - ax_i - bx_i^2)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow \sum_{i=1}^5 x_i y_i = a \sum_{i=1}^5 x_i^2 + b \sum_{i=1}^5 x_i^3$$

and  $\frac{\partial U}{\partial b} = 0$

$$\Rightarrow \sum_{i=1}^5 x_i^2 y_i = a \sum_{i=1}^5 x_i^3 + b \sum_{i=1}^5 x_i^4$$

Dropping the suffix  $i$ , Normal equations are

$$\Sigma xy = a \Sigma x^2 + b \Sigma x^3 \quad (24)$$

and  $\Sigma x^2 y = a \Sigma x^3 + b \Sigma x^4 \quad (25)$

Let us form a table as below:

$x$	$y$	$x^2$	$x^3$	$x^4$	$xy$	$x^2y$
1	1.8	1	1	1	1.8	1.8
2	5.1	4	8	16	10.2	20.4
3	8.9	9	27	81	26.7	80.1
4	14.1	16	64	256	56.4	225.6
5	19.8	25	125	625	99	495
Total		$\Sigma x^2 = 55$	$\Sigma x^3 = 225$	$\Sigma x^4 = 979$	$\Sigma xy = 194.1$	$\Sigma x^2y = 822.9$

Substituting these values in equations (24) and (25), we get

$$194.1 = 55a + 225b$$

and  $822.9 = 225a + 979b$

$$\Rightarrow a = \frac{83.85}{55} \simeq 1.52$$

and  $b = \frac{317.4}{664} \simeq .49$

Hence the required parabolic curve is  $y = 1.52x + 0.49x^2$ .



**Example 6.** Fit the curve  $pv^\gamma = k$  to the following data:

$p \text{ (kg/cm}^2\text{)}$	0.5	1	1.5	2	2.5	3
$v \text{ (liters)}$	1620	1000	750	620	520	460

**Sol.**

$$pv^\gamma = k$$

$$v = \left(\frac{k}{p}\right)^{1/\gamma} = k^{1/\gamma} p^{-1/\gamma}$$

Taking log,  $\log v = \frac{1}{\gamma} \log k - \frac{1}{\gamma} \log p$

which is of the form  $Y = A + BX$

where  $Y = \log v$ ,  $X = \log p$ ,  $A = \frac{1}{\gamma} \log k$  and  $B = -\frac{1}{\gamma}$

$p$	$v$	$X$	$Y$	$XY$	$X^2$
.5	1620	-.30103	3.20952	-.96616	0.09062
1	1000	0	3	0	0
1.5	750	.17609	2.87506	.50627	.03101
2	620	.30103	2.79239	.84059	.09062
2.5	520	.39794	2.716	1.08080	.15836
3	460	.47712	2.66276	1.27046	.22764
Total		$\Sigma X = 1.05115$	$\Sigma Y = 17.25573$	$\Sigma XY = 2.73196$	$\Sigma X^2 = .59825$

Here  $m = 6$

Normal equations are

$$17.25573 = 6A + 1.05115 B$$

and  $2.73196 = 1.05115 A + 0.59825 B$

Solving these, we get

$$A = 2.99911 \quad \text{and} \quad B = -0.70298$$

$$\therefore \gamma = -\frac{1}{B} = \frac{1}{.70298} = 1.42252$$

Again,  $\log k = \gamma A = 4.26629$

$$\therefore k = \text{antilog}(4.26629) = 18462.48$$

Hence the required curve is

$$pv^{1.42252} = 18462.48.$$

**Example 7.** Given the following experimental values:

$x$ :	0	1	2	3
$y$ :	2	4	10	15

Fit by the method of least squares a parabola of the type  $y = a + bx^2$ .

**Sol.** Error of estimate for  $i^{\text{th}}$  point  $(x_i, y_i)$  is  $E_i = (y_i - a - bx_i^2)$

By the principle of Least Squares, the values of  $a, b$  are such that

$$U = \sum_{i=1}^4 E_i^2 = \sum_{i=1}^4 (y_i - a - bx_i^2)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0 \Rightarrow \Sigma y = ma + b \Sigma x^2 \quad (26)$$

$$\text{and} \quad \frac{\partial U}{\partial b} = 0 \quad \Sigma x^2 y = a \Sigma x^2 + b \Sigma x^4 \quad (27)$$

$x$	$y$	$x^2$	$x^2 y$	$x^4$
0	2	0	0	0
1	4	1	4	1
2	10	4	40	16
3	15	9	135	81
Total	$\Sigma y = 31$	$\Sigma x^2 = 14$	$\Sigma x^2 y = 179$	$\Sigma x^4 = 98$

Here  $m = 4$

From (26) and (27),  $31 = 4a + 14b$  and  $179 = 14a + 98b$

Solving for  $a$  and  $b$ , we get  $a = 2.71, b = 1.44$

Hence the required curve is  $y = 2.71 + 1.44 x^2$ .

**Example 8.** The pressure of the gas corresponding to various volumes  $V$  is measured, given by the following data:

$V \text{ (cm}^3\text{):}$	50	60	70	90	100
$P \text{ (kg cm}^{-2}\text{):}$	64.7	51.3	40.5	25.9	78

Fit the data to the equation  $PV^\gamma = C$ .

**Sol.**  $PV^\gamma = C$

$$\Rightarrow P = CV^{-\gamma}$$

Take log on both sides,

$$\log P = \log C - \gamma \log V$$

$$\Rightarrow Y = A + BX$$

where  $Y = \log P$ ,  $A = \log C$ ,  $B = -\gamma$ ,  $X = \log V$

Normal equations are

$$\Sigma Y = mA + B\Sigma X$$

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

Here  $m = 5$

Table is as below:

$V$	$P$	$X = \log V$	$Y = \log P$	$XY$	$X^2$
50	64.7	1.69897	1.81090	3.07666	2.88650
60	51.3	1.77815	1.71012	3.04085	3.16182
70	40.5	1.84510	1.60746	2.96592	3.40439
90	25.9	1.95424	1.41330	2.76193	3.81905
100	78	2	1.89209	3.78418	4
		$\Sigma X = 9.27646$	$\Sigma Y = 8.43387$	$\Sigma XY = 15.62954$	$\Sigma X^2 = 17.27176$

From Normal equations, we have

$$8.43387 = 5A + 9.27646 B$$

and  $15.62954 = 9.27646 A + 17.27176 B$

Solving these, we get

$$A = 2.22476, B = -0.28997$$

$$\therefore \gamma = -B = 0.28997$$

$$C = \text{antilog}(A) = \text{antilog}(2.22476) = 167.78765$$

Hence the required equation of curve is

$$PV^{0.28997} = 167.78765.$$

independent variable. Either of the two may be estimated for the given values of the other. Thus, if we wish to estimate  $y$  for given values of  $x$ , we shall have the regression equation of the form  $y = a + bx$ , called the regression line of  $y$  on  $x$ . If we wish to estimate  $x$  for given values of  $y$ , we shall have the regression line of the form  $x = A + By$ , called the regression line of  $x$  on  $y$ .

Thus it implies, in general, *we always have two lines of regression.*

If the line of regression is so chosen that the sum of the squares of deviation parallel to the axis of  $y$  is minimized [See Figure (a)], it is called *the line of regression of  $y$  on  $x$*  and it gives *the best estimate of  $y$  for any given value of  $x$* .

If the line of regression is so chosen that the sum of the squares of deviations parallel to the axis of  $x$  is minimized [See Figure (b)], it is called *the line of regression of  $x$  on  $y$*  and it gives *the best estimate of  $x$  for any given value of  $y$* .

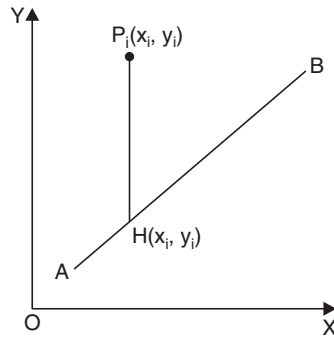


FIGURE (a)

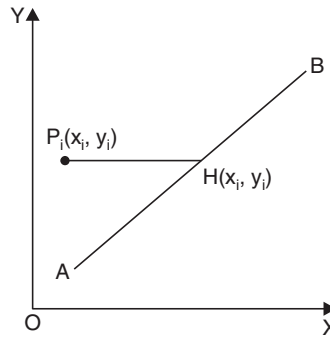


FIGURE (b)

The independent variable is called the *predictor* or *Regresser* or *Explanator* and the dependent variable is called the *predictant* or *Regressed* or *Explained* variable.

## 7.51 DERIVATION OF LINES OF REGRESSION

### 7.51.1 Line of Regression of $y$ on $x$

To obtain the line of regression of  $y$  on  $x$ , we shall assume  $y$  as dependent variable and  $x$  as independent variable.

Let  $y = a + bx$  be the equation of regression line of  $y$  on  $x$ .

The residual for  $i^{\text{th}}$  point is  $E_i = y_i - a - bx_i$ .

Introduce a new quantity  $U$  such that

$$U = \sum_{i=1}^n E_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 \quad (86)$$

According to the *principle of Least squares*, the constants  $a$  and  $b$  are chosen in such a way that the sum of the squares of residuals is minimum.

Now, the condition for  $U$  to be maximum or minimum is

$$\frac{\partial U}{\partial a} = 0 \quad \text{and} \quad \frac{\partial U}{\partial b} = 0$$

From (86), 
$$\frac{\partial U}{\partial a} = 2 \sum_{i=1}^n (y_i - a - bx_i)(-1)$$

$$\frac{\partial U}{\partial a} = 0 \text{ gives } 2 \sum_{i=1}^n (y_i - a - bx_i)(-1) = 0$$

$$\Rightarrow \boxed{\Sigma y = na + b \Sigma x} \quad (87)$$

Also, 
$$\frac{\partial U}{\partial b} = 2 \sum_{i=1}^n (y_i - a - bx_i)(-x_i)$$

$$\frac{\partial U}{\partial b} = 0 \text{ gives } 2 \sum_{i=1}^n (y_i - a - bx_i)(-x_i) = 0$$

$$\Rightarrow \boxed{\Sigma xy = a \Sigma x + b \Sigma x^2} \quad (88)$$

Equations (87) and (88) are called *normal equations*.

Solving (87) and (88) for ' $a$ ' and ' $b$ ', we get

$$b = \frac{\Sigma xy - \frac{1}{n} \Sigma x \Sigma y}{\Sigma x^2 - \frac{1}{n} (\Sigma x)^2} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} \quad (89)$$

and 
$$a = \frac{\Sigma y}{n} - b \frac{\Sigma x}{n} = \bar{y} - b\bar{x} \quad (90)$$

Eqn. (90) gives  $\bar{y} = a + b\bar{x}$

Hence  $y = a + bx$  line passes through point  $(\bar{x}, \bar{y})$ .

Putting  $a = \bar{y} - b\bar{x}$  in equation of line  $y = a + bx$ , we get

$$\boxed{y - \bar{y} = b(x - \bar{x})} \quad (91)$$

Equation (91) is called regression line of  $y$  on  $x$ . ' $b$ ' is called the regression coefficient of  $y$  on  $x$  and is usually denoted by  $b_{yx}$ .

Hence eqn. (91) can be rewritten as

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

where  $\bar{x}$  and  $\bar{y}$  are mean values while

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

In equation (88), shifting the origin to  $(\bar{x}, \bar{y})$ , we get

$$\sum (x - \bar{x})(y - \bar{y}) = a \sum (x - \bar{x}) + b \sum (x - \bar{x})^2$$

$$\Rightarrow nr \sigma_x \sigma_y = a(0) + bn \sigma_x^2$$

$$\begin{aligned} &\because \sum (x - \bar{x}) = 0 \\ &\frac{1}{n} \sum (x - \bar{x})^2 = \sigma_x^2 \\ &\text{and } \frac{\sum (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y} = r \end{aligned}$$

$$\Rightarrow b = r \frac{\sigma_y}{\sigma_x}$$

Hence regression coefficient  $b_{yx}$  can also be defined as

$$b_{yx} = r \frac{\sigma_y}{\sigma_x}$$

where  $r$  is the coefficient of correlation,  $\sigma_x$  and  $\sigma_y$  are the standard deviations of  $x$  and  $y$  series respectively.

### 7.51.2 Line of Regression of $x$ on $y$

Proceeding in the same way as 7.16.1, we can derive the regression line of  $x$  on  $y$  as

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

where  $b_{xy}$  is the regression coefficient of  $x$  on  $y$  and is given by

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

or

$$b_{xy} = r \frac{\sigma_x}{\sigma_y}$$

where the terms have their usual meanings.

NOTE

If  $r = 0$ , the two lines of regression become  $y = \bar{y}$  and  $x = \bar{x}$  which are two straight lines parallel to  $x$  and  $y$  axes respectively and passing through their means  $\bar{y}$  and  $\bar{x}$ . They are mutually perpendicular. If  $r = \pm 1$ , the two lines of regression will coincide.

## 7.52 USE OF REGRESSION ANALYSIS

---

- (i) In the field of Business, this tool of statistical analysis is widely used. Businessmen are interested in predicting future production, consumption, investment, prices, profits and sales etc.
- (ii) In the field of economic planning and sociological studies, projections of population, birth rates, death rates and other similar variables are of great use.

## 7.53 COMPARISON OF CORRELATION AND REGRESSION ANALYSIS

---

Both the correlation and regression analysis helps us in studying the relationship between two variables yet they differ in their approach and objectives.

- (i) Correlation studies are meant for studying the covariation of the two variables. They tell us whether the variables under study move in the same direction or in reverse directions. The degree of their covariation is also reflected in the correlation co-efficient but the correlation study does not provide the nature of relationship. It does not tell us about the relative movement in the variables and we cannot predict the value of one variable corresponding to the value of other variable. This is possible through regression analysis.
- (ii) Regression presumes one variable as a cause and the other as its effect. The independent variable is supposed to be affecting the dependent variable and as such we can estimate the values of the dependent variable by projecting the relationship between them. However, correlation between two series is not necessarily a cause-effect relationship.
- (iii) Coefficient of correlation cannot exceed unity but one of the regression coefficients can have a value higher than unity but the product of two regression coefficients can never exceed unity.

## 7.54 PROPERTIES OF REGRESSION CO-EFFICIENTS

**Property I. Correlation co-efficient is the geometric mean between the regression co-efficients.**

**Proof.** The co-efficients of regression are  $\frac{r\sigma_y}{\sigma_x}$  and  $\frac{r\sigma_x}{\sigma_y}$ .

Geometric mean between them =  $\sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r = \text{co-efficient of}$

correlation.

**Property II. If one of the regression co-efficients is greater than unity, the other must be less than unity.**

**Proof.** The two regression co-efficients are  $b_{yx} = \frac{r\sigma_y}{\sigma_x}$  and  $b_{xy} = \frac{r\sigma_x}{\sigma_y}$ .

Let  $b_{yx} > 1$ , then  $\frac{1}{b_{yx}} < 1$  (92)

Since  $b_{yx} \cdot b_{xy} = r^2 \leq 1$  ( $\because -1 \leq r \leq 1$ )

$\therefore b_{xy} \leq \frac{1}{b_{yx}} < 1$ . | using (92)

Similarly, if  $b_{xy} > 1$ , then  $b_{yx} < 1$ .

**Property III. Arithmetic mean of regression co-efficients is greater than the correlation co-efficient.**

**Proof.** We have to prove that

$$\frac{b_{yx} + b_{xy}}{2} > r$$

or  $r \frac{\sigma_y}{\sigma_x} + r \frac{\sigma_x}{\sigma_y} > 2r$

or  $\sigma_x^2 + \sigma_y^2 > 2\sigma_x\sigma_y$

or  $(\sigma_x - \sigma_y)^2 > 0$  which is true.

**Property IV. Regression co-efficients are independent of the origin but not of scale.**

**Proof.** Let  $u = \frac{x-a}{h}$ ,  $v = \frac{y-b}{k}$ , where  $a$ ,  $b$ ,  $h$  and  $k$  are constants

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r \cdot \frac{k\sigma_v}{h\sigma_u} = \frac{k}{h} \left( \frac{r\sigma_v}{\sigma_u} \right) = \frac{k}{h} b_{vu}$$



Similarly,  $b_{xy} = \frac{h}{k} b_{uv}$ .

Thus,  $b_{yx}$  and  $b_{xy}$  are both independent of  $a$  and  $b$  but not of  $h$  and  $k$ .

**Property V. The correlation co-efficient and the two regression co-efficients have same sign.**

**Proof.** Regression co-efficient of  $y$  on  $x = b_{yx} = r \frac{\sigma_y}{\sigma_x}$

Regression co-efficient of  $x$  on  $y = b_{xy} = r \frac{\sigma_x}{\sigma_y}$

Since  $\sigma_x$  and  $\sigma_y$  are both positive;  $b_{yx}$ ,  $b_{xy}$  and  $r$  have same sign.

## 7.55 ANGLE BETWEEN TWO LINES OF REGRESSION

If  $\theta$  is the acute angle between the two regression lines in the case of two variables  $x$  and  $y$ , show that

$$\tan \theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}, \quad \text{where } r, \sigma_x, \sigma_y \text{ have their usual meanings.}$$

Explain the significance of the formula when  $r = 0$  and  $r = \pm 1$ .

**Proof.** Equations to the lines of regression of  $y$  on  $x$  and  $x$  on  $y$  are

$$y - \bar{y} = \frac{r\sigma_y}{\sigma_x} (x - \bar{x}) \quad \text{and} \quad x - \bar{x} = \frac{r\sigma_x}{\sigma_y} (y - \bar{y})$$

Their slopes are  $m_1 = \frac{r\sigma_y}{\sigma_x}$  and  $m_2 = \frac{\sigma_y}{r\sigma_x}$ .

$$\begin{aligned} \therefore \tan \theta &= \pm \frac{m_2 - m_1}{1 + m_2 m_1} = \pm \frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y}{\sigma_x} \cdot \frac{r\sigma_y}{\sigma_x}} \\ &= \pm \frac{1-r^2}{r} \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \pm \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \end{aligned}$$

Since  $r^2 \leq 1$  and  $\sigma_x, \sigma_y$  are positive.

$\therefore$  +ve sign gives the acute angle between the lines.

Hence 
$$\tan \theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

when  $r = 0$ ,  $\theta = \frac{\pi}{2}$   $\therefore$  The two lines of regression are perpendicular to each other.

Hence the estimated value of  $y$  is the same for all values of  $x$  and vice-versa.

When  $r = \pm 1$ ,  $\tan \theta = 0$  so that  $\theta = 0$  or  $\pi$

Hence the lines of regression coincide and there is perfect correlation between the two variates  $x$  and  $y$ .

## 7.56 ALGORITHM FOR LINEAR REGRESSION

---

1. Read  $n$
2.  $\text{sum } x \leftarrow 0$
3.  $\text{sum } xsq \leftarrow 0$
4.  $\text{sum } y \leftarrow 0$
5.  $\text{sum } xy \leftarrow 0$
6. for  $i = 1$  to  $n$  do
7. Read  $x, y$
8.  $\text{sum } x \leftarrow \text{sum } x + x$
9.  $\text{sum } xsq \leftarrow \text{sum } xsq + x^2$
10.  $\text{sum } y \leftarrow \text{sum } y + y$
11.  $\text{sum } xy \leftarrow \text{sum } xy + x \times y$
- end for
12.  $\text{denom} \leftarrow n \times \text{sum } x \text{ sq} - \text{sum } x \times \text{sum } x$
13.  $a \leftarrow (\text{sum } y \times \text{sum } x \text{ sq} - \text{sum } x \times \text{sum } xy) / \text{denom}$
14.  $b \leftarrow (n \times \text{sum } xy - \text{sum } x \times \text{sum } y) / \text{denom}$
15. Write  $b, a$
16. Stop

### 7.57 PROGRAM TO IMPLEMENT LEAST SQUARE FIT OF A REGRESSION LINE OF Y ON X

---

```

#include<stdio.h>
#include<conio.h>
#include<math.h>
void main()
{
    int data,i;
    float x[10],y[10],xy[10],x2[10],z;
    float sum1=0.0,sum2=0.0,sum3=0.0,sum4=0.0;
    clrscr();
    printf("Enter the number of data points:");
    scanf("%d",&data);
    printf("Enter the value of x: \n");
    for(i=0;i<data;i++)
    {
        printf("Value of x%d:",i+1);
        scanf("%f",&x[i]);
    }
    printf("\nEnter the value of f(x):\n");
    for(i=0;i<data;i++)
    {
        printf("Value of f(x%d):",i+1);
        scanf("%f",&y[i]);
    }
    for(i=0;i<data;i++)
    {
        xy[i]=x[i]*y[i];
        x2[i]=x[i]*x[i];
        sum1 +=xy[i];
        sum2 +=x2[i];
        sum3 +=x[i];
        sum4 +=y[i];
    }
}

```

```

sum3 =sum3/2;
sum4 =sum4/2;
//printf("%.2f %.2f %.2f", %.2f" sum1,sum2,sum3,sum4);
sum1=(sum1/sum2);
z=(sum1*sum3)-sum4;
printf("\n\nThe REGRESSION LINE OF Y on X is:\n");
printf("\t\t\t\t y=%.2f *x - (%.2f)",sum1,z);
getch(1);
}

```

### 7.58 PROGRAM TO IMPLEMENT LEAST SQUARE FIT OF A REGRESSION LINE OF X ON Y

---

```

#include<stdio.h>
#include<conio.h>
#include<math.h>
void main()
{
    int data,i;
    float x[10],y[10],xy[10],y2[10],z;
    float sumx=0.0,sumy=0.0,sumxy=0.0,sumy2=0.0;
    clrscr();
    printf("Enter the number of data points: ");
    scanf("%d",&data);
    printf("Enter the value of x: \n");
    for(i=0;i<data;i++)
    {
        printf("Value of x%d: ",i+1);
        scanf("%f",&x[i]);
    }
    printf("\nEnter the value of f(x): \n");
    for(i=0;i<data; i++)
    {
        printf("Value of f(x%d):", i+1);
        scanf("%f",&y[i]);
    }
}

```

```

for(i=0;i<data;i++)
{
    xy[i]=x[i]*y[i];
    y2[i]=y[i]*y[i];
    sumxy +=xy[i];
    sumy2 +=y2[i];
    sumx +=x[i];
    sumy +=y[i];
}
sumx =sumx/2;
sumy =sumy/2;
sumxy=(sumxy/sumy2);
z=(sumxy*sumy)-sumx;
printf("\n\nThe REGRESSION LINE OF X on Y is:\n");
printf("\t\t\t x = %.2f *y - (%.2f)",sumxy, z);
getch();
}

```

### EXAMPLES

**Example 1.** If the regression coefficients are 0.8 and 0.2, what would be the value of coefficient of correlation?

**Sol.** We know that,

$$r^2 = b_{yx} \cdot b_{xy} = 0.8 \times 0.2 = 0.16$$

Since  $r$  has the same sign as both the regression coefficients  $b_{yx}$  and  $b_{xy}$

Hence  $r = \sqrt{0.16} = 0.4.$

$$r^2 = b_{yx} \cdot b_{xy}$$

**Example 2.** Calculate linear regression coefficients from the following:

$x$	$\rightarrow$	1	2	3	4	5	6	7	8
$y$	$\rightarrow$	3	7	10	12	14	17	20	24

**Sol.** Linear regression coefficients are given by

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

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

Let us prepare the following table:

$x$	$y$	$x^2$	$y^2$	$xy$
1	3	1	9	3
2	7	4	49	14
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
6	17	36	289	102
7	20	49	400	140
8	24	64	576	192
$\Sigma x = 36$	$\Sigma y = 107$	$\Sigma x^2 = 204$	$\Sigma y^2 = 1763$	$\Sigma xy = 599$

Here  $n = 8$

$$\therefore b_{yx} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 204) - (36)^2} = \frac{4792 - 3852}{1632 - 1296} = \frac{940}{336} = 2.7976$$

and 
$$b_{xy} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 1763) - (107)^2} = \frac{940}{2655} = 0.3540$$

**Example 3.** The following table gives age ( $x$ ) in years of cars and annual maintenance cost ( $y$ ) in hundred rupees:

$x$ :	1	3	5	7	9
$y$ :	15	18	21	23	22

Estimate the maintenance cost for a 4 year old car after finding the regression equation.

**Sol.**

$x$	$y$	$xy$	$x^2$
1	15	15	1
3	18	54	9
5	21	105	25
7	23	161	49
9	22	198	81
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma xy = 533$	$\Sigma x^2 = 165$

Here,  $n = 5$

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

$$\bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$$

$$\therefore b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(5 \times 533) - (25 \times 99)}{(5 \times 165) - (25)^2} = 0.95$$

Regression line of  $y$  on  $x$  is given by

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

$$\Rightarrow y - 19.8 = 0.95 (x - 5)$$

$$\Rightarrow y = 0.95x + 15.05$$

$$\text{When } x = 4 \text{ years, } y = (0.95 \times 4) + 15.05$$

$$= 18.85 \text{ hundred rupees} = \text{Rs. } 1885.$$

**Example 4.** In a partially destroyed laboratory record of an analysis of a correlation data, the following results only are eligible:

✓ Variance of  $x = 9$

Regression equations:  $8x - 10y + 66 = 0$ ,  $40x - 18y = 214$ .

What were (a) the mean values of  $x$  and  $y$  (b) the standard deviation of  $y$  and the co-efficient of correlation between  $x$  and  $y$ .

**Sol.** (a) Since both lines of regression pass through the point  $(\bar{x}, \bar{y})$  therefore, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \quad (93)$$

$$40\bar{x} - 18\bar{y} - 214 = 0 \quad (94)$$

$$\text{Multiplying (93) by 5, } 40\bar{x} - 50\bar{y} + 330 = 0 \quad (95)$$

$$\text{Subtracting (95) from (94), } 32\bar{y} - 544 = 0$$

$$\therefore \bar{y} = 17 \quad \checkmark$$

$$\therefore \text{ From (93), } 8\bar{x} - 170 + 66 = 0$$

$$\text{or } 8\bar{x} = 104 \quad \therefore \bar{x} = 13 \quad \checkmark$$

$$\text{Hence } \bar{x} = 13, \bar{y} = 17 \quad \checkmark$$

$$(b) \text{ Variance of } x = \sigma_x^2 = 9 \quad (\text{given})$$

$$\therefore \sigma_x = 3 \quad \checkmark$$

The equations of lines of regression can be written as

$$y = .8x + 6.6 \quad \text{and} \quad x = .45y + 5.35$$

$$\therefore \text{The regression co-efficient of } y \text{ on } x \text{ is } \frac{r\sigma_y}{\sigma_x} = .8 \quad (96)$$

$$\text{The regression co-efficient of } x \text{ on } y \text{ is } \frac{r\sigma_x}{\sigma_y} = .45 \quad (97)$$

$$\text{Multiplying (96) and (97), } r^2 = .8 \times .45 = .36 \quad \therefore r = 0.6$$

(+ve sign with square root is taken because regression co-efficients are +ve).

$$\text{From (96),} \quad \sigma_y = \frac{.8\sigma_x}{r} = \frac{.8 \times 3}{0.6} = 4.$$

**Example 5.** The regression lines of  $y$  on  $x$  and  $x$  on  $y$  are respectively  $y = ax + b$ ,  $x = cy + d$ . Show that

$$\frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}, \quad \bar{x} = \frac{bc + d}{1 - ac} \quad \text{and} \quad \bar{y} = \frac{ad + b}{1 - ac}.$$

**Sol.** The regression line of  $y$  on  $x$  is

$$y = ax + b \quad (98)$$

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

The regression line of  $x$  on  $y$  is

$$x = cy + d \quad (99)$$

$$\therefore b_{xy} = c$$

$$\text{We know that,} \quad b_{yx} = r \frac{\sigma_y}{\sigma_x} \quad (100)$$

$$\text{and} \quad b_{xy} = r \frac{\sigma_x}{\sigma_y} \quad (101)$$

Dividing eqn. (100) by (101), we get

$$\frac{b_{yx}}{b_{xy}} = \frac{\sigma_y^2}{\sigma_x^2} \Rightarrow \frac{a}{c} = \frac{\sigma_y^2}{\sigma_x^2} \Rightarrow \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}$$

Since both the regression lines pass through the point  $(\bar{x}, \bar{y})$  therefore,

$$\bar{y} = a\bar{x} + b \quad \text{and} \quad \bar{x} = c\bar{y} + d$$

$$\Rightarrow a\bar{x} - \bar{y} = -b \quad (102)$$

$$\bar{x} - c\bar{y} = d \quad (103)$$



Multiplying equation (103) by  $a$  and then subtracting from (102), we get

$$(ac - 1) \bar{y} = -ad - b \Rightarrow \bar{y} = \frac{ad + b}{1 - ac}$$

Similarly, we get  $\bar{x} = \frac{bc + d}{1 - ac}$ .

**Example 6.** For two random variables,  $x$  and  $y$  with the same mean, the two regression equations are

$$y = ax + b \quad \text{and} \quad x = \alpha y + \beta$$

Show that  $\frac{b}{\beta} = \frac{1 - \alpha}{1 - a}$ .

Find also the common mean.

**Sol.** Here,  $b_{yx} = a$ ,  $b_{xy} = \alpha$

Let the common mean be  $m$ , then regression lines are

$$\begin{aligned} y - m &= a(x - m) \\ \Rightarrow y &= ax + m(1 - a) \end{aligned} \tag{104}$$

and  $x - m = \alpha(y - m)$

$$\Rightarrow x = \alpha y + m(1 - \alpha) \tag{105}$$

Comparing (104) and (105) with the given equations.

$$\begin{aligned} b &= m(1 - a), \beta = m(1 - \alpha) \\ \therefore \frac{b}{\beta} &= \frac{1 - a}{1 - \alpha} \end{aligned}$$

Again  $m = \frac{b}{1 - a} = \frac{\beta}{1 - \alpha}$

Since regression lines pass through  $(\bar{x}, \bar{y})$

$$\therefore \bar{x} = \alpha \bar{y} + \beta$$

and  $\bar{y} = a\bar{x} + b$  will hold.

$$\Rightarrow m = am + b$$

$$m = \alpha m + \beta$$

$$\Rightarrow am + b = \alpha m + \beta$$

$$\Rightarrow m = \frac{\beta - b}{a - \alpha}$$

**Example 7.** Obtain the line of regression of  $y$  on  $x$  for the data given below:

$x$ :	1.53	1.78	2.60	2.95	3.42
$y$ :	33.50	36.30	40.00	45.80	53.50.

**Sol.** The line of regression of  $y$  on  $x$  is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) \quad (106)$$

where  $b_{yx}$  is the coefficient of regression given by

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

Now we form the table as,

$x$	$y$	$x^2$	$xy$
1.53	33.50	2.3409	51.255
1.78	36.30	2.1684	64.614
2.60	40.00	6.76	104
2.95	45.80	8.7025	135.11
3.42	53.50	11.6964	182.97
$\Sigma x = 12.28$	$\Sigma y = 209.1$	$\Sigma x^2 = 32.6682$	$\Sigma xy = 537.949$

Here,  $n = 5$

$$b_{yx} = \frac{(5 \times 537.949) - (12.28 \times 209.1)}{(5 \times 32.6682) - (12.28)^2} = \frac{121.997}{12.543} = 9.726$$

Also,  $\text{mean } \bar{x} = \frac{\Sigma x}{n} = \frac{12.28}{5} = 2.456$

and  $\bar{y} = \frac{\Sigma y}{n} = \frac{209.1}{5} = 41.82$

$\therefore$  From (106), we get

$$y - 41.82 = 9.726(x - 2.456) = 9.726x - 23.887$$

$$y = 17.932 + 9.726x$$

which is the required line of regression of  $y$  on  $x$ .

**Example 8.** For 10 observations on price ( $x$ ) and supply ( $y$ ), the following data were obtained (in appropriate units):

$$\Sigma x = 130, \quad \Sigma y = 220, \quad \Sigma x^2 = 2288, \quad \Sigma y^2 = 5506 \text{ and } \Sigma xy = 3467$$

Obtain the two lines of regression and estimate the supply when the price is 16 units.

**Sol.** Here,  $n = 10$ ,  $\bar{x} = \frac{\Sigma x}{n} = 13$  and  $\bar{y} = \frac{\Sigma y}{n} = 22$

Regression coefficient of  $y$  on  $x$  is

$$\begin{aligned} b_{yx} &= \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 2288) - (130)^2} \\ &= \frac{34670 - 28600}{22880 - 16900} = \frac{6070}{5980} = 1.015 \end{aligned}$$

$\therefore$  Regression line of  $y$  on  $x$  is

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

$$y - 22 = 1.015(x - 13)$$

$$\Rightarrow y = 1.015x + 8.805$$

Regression coefficient of  $x$  on  $y$  is

$$\begin{aligned} b_{xy} &= \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma y^2 - (\Sigma y)^2} \\ &= \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 5506) - (220)^2} = \frac{6070}{6660} = 0.9114 \end{aligned}$$

Regression line of  $x$  on  $y$  is

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

$$x - 13 = 0.9114(y - 22)$$

$$x = 0.9114y - 7.0508$$

Since we are to estimate supply ( $y$ ) when price ( $x$ ) is given therefore we are to use regression line of  $y$  on  $x$  here.

When  $x = 16$  units,

$$y = 1.015(16) + 8.805 = 25.045 \text{ units.}$$

**Example 9.** The following results were obtained from records of age ( $x$ ) and systolic blood pressure ( $y$ ) of a group of 10 men:

	$x$	$y$
Mean	53	142
Variance	130	165

and  $\Sigma(x - \bar{x})(y - \bar{y}) = 1220$

Find the approximate regression equation and use it to estimate the blood pressure of a man whose age is 45.

**Sol.** Given:

Mean	$\bar{x} = 53$
Mean	$\bar{y} = 142$
Variance	$\sigma_x^2 = 130$
Variance	$\sigma_y^2 = 165$
Number of men,	$n = 10$

$$\Sigma(x - \bar{x})(y - \bar{y}) = 1220$$

$\therefore$  Coefficient of correlation,

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y} = \frac{1220}{10\sqrt{130 \times 165}} = \frac{122}{146.458} = 0.83.$$

Since we are to estimate blood pressure ( $y$ ) of a 45 years old man, we will find regression line of  $y$  on  $x$ .

$$\text{Regression coefficient } b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.83 \times \sqrt{\frac{165}{130}} = 0.935.$$

Regression line of  $y$  on  $x$  is given by

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

$$\Rightarrow y - 142 = 0.935(x - 53) = 0.935x - 49.555$$

$$\Rightarrow y = 0.935x + 92.445$$

when  $x = 45$ ,

$$y = (0.935 \times 45) + 92.445 = 134.52.$$

Hence the required blood pressure = 134.52.

**Example 10.** The following results were obtained from scores in Applied Mechanics and Engineering Mathematics in an examination:

	Applied Mechanics ( $x$ )	Engineering Mathematics ( $y$ )
Mean	47.5	39.5
Standard Deviation	16.8	10.8

$$r = 0.95.$$

Find both the regression equations. Also estimate the value of  $y$  for  $x = 30$ .

**Sol.**  $\bar{x} = 47.5,$   $\bar{y} = 39.5$   
 $\sigma_x = 16.8,$   $\sigma_y = 10.8$  and  $r = 0.95.$

Regression coefficients are

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = 0.95 \times \frac{10.8}{16.8} = 0.6107$$

and 
$$b_{xy} = r \frac{\sigma_x}{\sigma_y} = 0.95 \times \frac{16.8}{10.8} = 1.477.$$

Regression line of  $y$  on  $x$  is

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

$$\begin{aligned} \Rightarrow y - 39.5 &= 0.6107(x - 47.5) = 0.6107x - 29.008 \\ y &= 0.6107x + 10.49 \end{aligned} \quad (107)$$

Regression line of  $x$  on  $y$  is

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

$$\begin{aligned} \Rightarrow x - 47.5 &= 1.477(y - 39.5) \\ \Rightarrow x - 47.5 &= 1.477y - 58.3415 \\ x &= 1.477y - 10.8415 \end{aligned}$$

Putting  $x = 30$  in equation (107), we get

$$y = (0.6107)(30) + 10.49 = 18.321 + 10.49 = 28.81.$$

**Example 11.** From the following data. Find the most likely value of  $y$  when  $x = 24$ :

	$y$	$x$
Mean	985.8	18.1
S.D.	36.4	2.0

$$r = 0.58.$$

**Sol.** Given:  $\bar{y} = 985.8$ ,  $\bar{x} = 18.1$ ,  $\sigma_y = 36.4$ ,  $\sigma_x = 2$ ,  $r = 0.58$

Regression coefficient,

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} = (0.58) \frac{36.4}{2} = 10.556.$$

Regression line of  $y$  on  $x$  is

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

$$\begin{aligned} \Rightarrow y - 985.8 &= 10.556(x - 18.1) \\ y - 985.8 &= 10.556x - 191.06 \end{aligned}$$

$\Rightarrow$   $y = 10.556x + 794.73$   
 when  $x = 24$ ,

$$y = (10.556 \times 24) + 794.73$$

$$y = 1048 \text{ (approximately).}$$

**Example 12.** The equations of two regression lines, obtained in a correlation analysis of 60 observations are:

$$5x = 6y + 24 \text{ and } 1000y = 768x - 3608.$$

What is the correlation coefficient? Show that the ratio of coefficient of variability of  $x$  to that of  $y$  is  $\frac{5}{24}$ . What is the ratio of variances of  $x$  and  $y$ ?

**Sol.** Regression line of  $x$  on  $y$  is

$$5x = 6y + 24$$

$$x = \frac{6}{5}y + \frac{24}{5}$$

$$\therefore b_{xy} = \frac{6}{5} \quad (108)$$

Regression line of  $y$  on  $x$  is

$$1000y = 768x - 3608$$

$$y = 0.768x - 3.608$$

$$\therefore b_{yx} = 0.768 \quad (109)$$

$$\text{From (108), } r \frac{\sigma_x}{\sigma_y} = \frac{6}{5} \quad (110)$$

$$\text{From (109), } r \frac{\sigma_y}{\sigma_x} = 0.768 \quad (111)$$

Multiplying equations (110) and (111), we get

$$r^2 = 0.9216 \Rightarrow r = 0.96 \quad (112)$$

Dividing (111) by (110), we get

$$\frac{\sigma_x^2}{\sigma_y^2} = \frac{6}{5 \times 0.768} = 1.5625.$$

Taking the square root, we get

$$\frac{\sigma_x}{\sigma_y} = 1.25 = \frac{5}{4} \quad (113)$$

Since the regression lines pass through the point  $(\bar{x}, \bar{y})$ , we have

$$5\bar{x} = 6\bar{y} + 24$$

$$1000\bar{y} = 768\bar{x} - 3608.$$

Solving the above equations for  $\bar{x}$  and  $\bar{y}$ , we get

$$\bar{x} = 6, \bar{y} = 1.$$

Coefficient of variability of  $x = \frac{\sigma_x}{\bar{x}},$

Coefficient of variability of  $y = \frac{\sigma_y}{\bar{y}}.$

$$\therefore \text{ Required ratio} = \frac{\sigma_x}{\bar{x}} \times \frac{\bar{y}}{\sigma_y} = \frac{\bar{y}}{\bar{x}} \left( \frac{\sigma_x}{\sigma_y} \right) = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}. \quad | \text{ using (113)}$$

**Example 13.** The following data regarding the heights ( $y$ ) and weights ( $x$ ) of 100 college students are given:

$$\Sigma x = 15000, \Sigma x^2 = 2272500, \Sigma y = 6800, \Sigma y^2 = 463025 \text{ and } \Sigma xy = 1022250.$$

Find the equation of the regression line of height on weight.

**Sol.** 
$$\bar{x} = \frac{\Sigma x}{n} = \frac{15000}{100} = 150$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{6800}{100} = 68$$

Regression coefficient of  $y$  on  $x$ ,

$$\begin{aligned} b_{yx} &= \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(100 \times 1022250) - (15000 \times 6800)}{(100 \times 2272500) - (15000)^2} \\ &= \frac{102225000 - 102000000}{227250000 - 225000000} \\ &= \frac{225000}{2250000} = 0.1 \end{aligned}$$

Regression line of height ( $y$ ) on weight ( $x$ ) is given by

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

$$\Rightarrow y - 68 = 0.1(x - 150)$$

$$\Rightarrow y = 0.1x - 15 + 68$$

$$\Rightarrow y = 0.1x + 53.$$

**Example 14.** Find the coefficient of correlation when the two regression equations are

$$X = -0.2Y + 4.2$$

$$Y = -0.8X + 8.4.$$

**Sol.** We have the regression lines

$$X = -0.2Y + 4.2 \quad (114)$$

$$Y = -0.8X + 8.4. \quad (115)$$

Let us assume that eqn. (114) is the regression line of X on Y and eqn. (115) is the regression line of Y on X then,

Regression coefficient of X on Y is

$$b_{XY} = -0.2$$

Regression coefficient of Y on X is

$$b_{YX} = -0.8$$

Since  $b_{XY}$  and  $b_{YX}$  are of the same sign and  $b_{XY}b_{YX} = 0.16 (< 1)$  hence our assumption is correct.

We know that

$$b_{XY}b_{YX} = r^2 \quad | \text{ where } r \text{ is the correlation coefficient}$$

$$\Rightarrow (-0.2)(-0.8) = r^2$$

$$\Rightarrow r^2 = 0.16$$

$$\Rightarrow r = -0.4. \quad | \text{ Since } r, \sigma_x \text{ and } \sigma_y \text{ have the same sign}$$

**Example 15.** A panel of two judges, A and B, graded seven TV serial performances by awarding scores independently as shown in the following table:

Performance	1	2	3	4	5	6	7
Scores by A	46	42	44	40	43	41	45
Scores by B	40	38	36	35	39	37	41

The eighth TV performance, which judge B could not attend, was awarded 37 scores by judge A. If judge B had also been present, how many scores would be expected to have been awarded by him to the eighth TV performance?

Use regression analysis to answer this question.

**Sol.** Let the scores awarded by judge A be denoted by  $x$  and the scores awarded by judge B be denoted by  $y$ .



Here,  $n = 7$ ;  $\bar{x} = \frac{\Sigma x}{n} = \frac{46 + 42 + 44 + 40 + 43 + 41 + 45}{7} = 43$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{40 + 38 + 36 + 35 + 39 + 37 + 41}{7} = 38$$

Let us form the table as

$x$	$y$	$xy$	$x^2$
46	40	1840	2116
42	38	1596	1764
44	36	1584	1936
40	35	1400	1600
43	39	1677	1849
41	37	1517	1681
45	41	1845	2025
$\Sigma x = 301$	$\Sigma y = 266$	$\Sigma xy = 11459$	$\Sigma x^2 = 12971$

Regression coefficient,

$$b_{yx} = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2} = \frac{(7 \times 11459) - (301 \times 266)}{(7 \times 12971) - (301)^2}$$

$$= \frac{80213 - 80066}{90797 - 90601} = \frac{147}{196} = 0.75$$

Regression line of  $y$  on  $x$  is given by

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

$$y - 38 = 0.75(x - 43)$$

$$\Rightarrow y = 0.75x + 5.75$$

when  $x = 37$ ,

$$y = 0.75(37) + 5.75 = 33.5 \text{ marks}$$

Hence, if judge B had also been present, 33.5 scores would be expected to have been awarded to the eighth T.V. performance.

### ASSIGNMENT 7.5

- Find the regression line of  $y$  on  $x$  from the following data:

$x$ :	1	2	3	4	5
$y$ :	2	5	3	8	7