

The term correlation refers to the relationship between the variables. Simple correlation refers to the relationship between two variables. There may be fluctuation or covariation is to find how well there exists a linear relationship between two variables. The direction of change and the closeness of the relationship are found.

Various **types of correlation** are considered under the following three heads.

Positive or negative. When the values of two variables change in the same direction, there is **positive correlation** between the two variables.

| | | | | | | | |
|-------------|---|----|----|----|----|-----|-----|
| Example 1 : | X | 50 | 60 | 70 | 95 | 100 | 105 |
| | Y | 23 | 32 | 37 | 41 | 46 | 50 |
| Example 2 : | X | 34 | 25 | 18 | 10 | 7 | |
| | Y | 51 | 49 | 42 | 33 | 19 | |

In the two examples, X and Y change in the same direction (X and Y increase in Ex. 1 and they decrease in Ex.2). Hence, there is positive correlation. Positive correlation is generally found between the following pairs of variables.

1. Price and Supply.
2. Sales and Expenditure on Advertisement.
3. Yield and Fertiliser Applied.

When the values of two variables change in the opposite directions, there is **negative correlation** between the two variables.

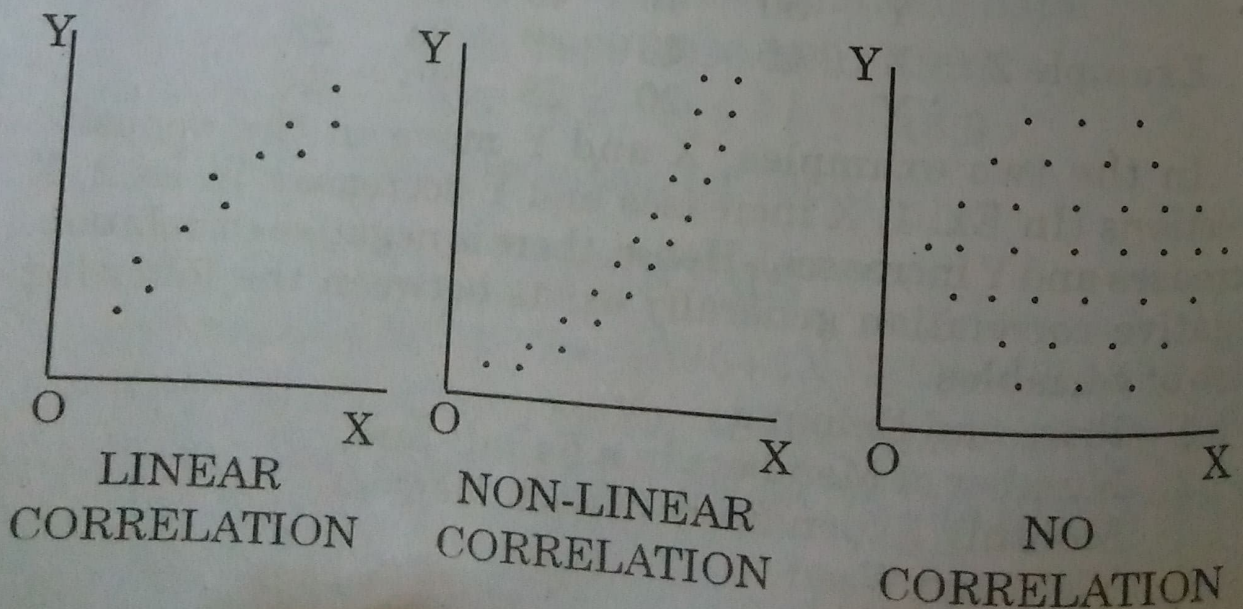
| | | | | | | | |
|-------------|---|----|----|----|----|-----|-----|
| Example 1 : | X | 50 | 60 | 70 | 95 | 100 | 105 |
| | Y | 50 | 46 | 40 | 30 | 24 | 15 |
| Example 2 : | X | 45 | 43 | 39 | 34 | 28 | |
| | Y | 14 | 20 | 28 | 29 | 34 | |

In the two examples, X and Y move in the opposite directions (In Ex. 1, X increases and Y decreases; in Ex.2, X decreases and Y increases). Hence, there is negative correlation. Negative correlation generally exists between the following pairs of variables.

1. Price and Demand
2. Number of Members in a Family and Monthly Expenditure of each Member.
3. Yield and Weed

Simple or Partial or Multiple. When only two variables are considered as under positive or negative correlation above, the correlation between them is called **simple correlation**. When more than two variables are considered, the correlation between two of them when all other variables are held constant, i.e., when the linear effects of all other variables on them are removed, is called **partial correlation**. When more than two variables are considered, the correlation between one of them and its estimate based on the group consisting of the other variables is called **multiple correlation**. In Agriculture, yield, rainfall and fertiliser are interrelated. Similarly, in Economics, price, demand and income are interrelated. The correlation between price and demand when the effect of income is removed is (an example for) partial correlation. The correlation between demand and the estimate of demand as given by the group of variables, price and income is (an example for) multiple correlation.

Linear or Non-linear or No Correlation. Corresponding to each pair of values of two variables, plot a point on a graph sheet. Consider all the points so obtained. When all the points lie exactly on a line or scattered around a line, there is linear correlation between the two variables. When all the points lie exactly on a curve or scattered around a curve, there is non-linear correlation between the two variables. When the points are scattered neither around a line nor around a curve, there is no correlation between the two variables. The following diagrams depict these three kinds.



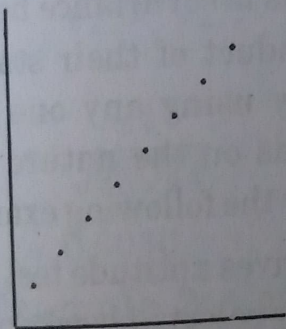
Methods. The following four methods are available under simple linear correlation and among them, product moment method is the best one.

- (i) Scatter Diagram
- (ii) Karl Pearson's correlation coefficient or product moment correlation coefficient (r)
- (iii) Spearman's rank correlation coefficient (ρ)
- (iv) Correlation coefficient by concurrent deviation method (r_c)

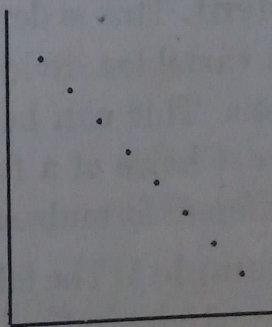
SCATTER DIAGRAM

Let (x_i, y_i) $i = 1, 2, 3 \dots N$ be the pairs of values of two variables X and Y. A point is plotted on a graph sheet corresponding to each pair of the values. The resulting diagram with N points is called **scatter diagram**.

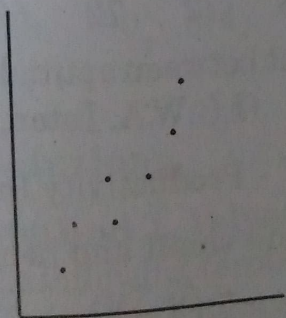
Possible types of scatter diagrams under simple linear correlation are as given below. From a diagram, it can be found out whether the correlation is positive or negative and whether it is perfect or high or low.



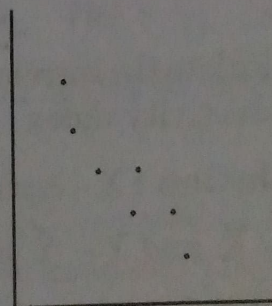
PERFECT POSITIVE CORRELATION



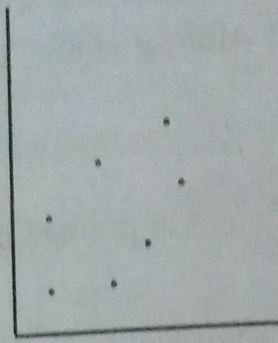
PERFECT NEGATIVE CORRELATION



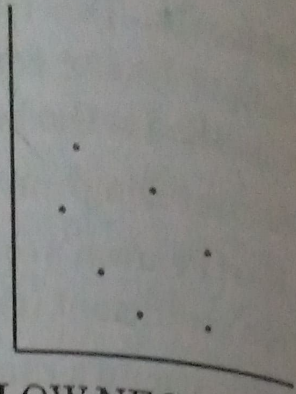
HIGH POSITIVE CORRELATION



HIGH NEGATIVE CORRELATION



LOW POSITIVE
CORRELATION



LOW NEGATIVE
CORRELATION

The merits of this method are as follows. This is easy to draw, non mathematical and simple to understand. This does not involve computations. The greatest demerit is that this is not quantitative. As no numerical value is computed, comparison is not possible sometimes. Decisions based on this are not as accurate as those based on correlation coefficients.

KARL PEARSON'S COEFFICIENT OF CORRELATION (r)

This is also called **product moment correlation coefficient**. This is denoted by r . This is covariance between the two variables divided by the product of their standard deviations. This can be calculated by using any one of the formulae. Choice of a formula depends on the nature of the data. Different formulae are seen under the following examples.

Example 1: The following table gives aptitude test scores and productivity indices of 8 randomly selected workers:

| | | | | | | | | |
|----------------------|----|----|----|----|----|----|----|----|
| Aptitude Score : | 57 | 58 | 59 | 59 | 60 | 61 | 62 | 64 |
| Productivity Index : | 67 | 68 | 65 | 68 | 72 | 72 | 69 | 71 |

Calculate the correlation coefficient between aptitude score and productivity index.

(I.C.W.A. Inter., D 92)

Solution : X - Aptitude Score; Y - Productivity Index.

$X - \bar{X}$ and $Y - \bar{Y}$ are integers and small and hence the following formula is used. $\sum x = \sum (X - \bar{X}) = 0$ and $\sum y = \sum (Y - \bar{Y}) = 0$ are the properties.

| X | Y | $x = X - \bar{X}$ $\bar{X} = 60$ | $y = Y - \bar{Y}$ $\bar{Y} = 69$ | xy | x^2 | y^2 |
|---------------------|---------------------|-------------------------------------|-------------------------------------|---------------------|----------------------|----------------------|
| 57 | 67 | -3 | -2 | 6 | 9 | 4 |
| 58 | 68 | -2 | -1 | 2 | 4 | 1 |
| 59 | 65 | -1 | -4 | 4 | 1 | 16 |
| 59 | 68 | -1 | -1 | 1 | 1 | 1 |
| 60 | 72 | 0 | 3 | 0 | 0 | 9 |
| 61 | 72 | 1 | 3 | 3 | 1 | 9 |
| 62 | 69 | 2 | 0 | 0 | 4 | 0 |
| 62 | 71 | 4 | 2 | 8 | 16 | 4 |
| 64 | 71 | 4 | 2 | 8 | 16 | 4 |
| $\Sigma X =$ 480 | $\Sigma Y =$ 552 | $\Sigma x =$ 0 | $\Sigma y =$ 0 | $\Sigma xy =$ 24 | $\Sigma x^2 =$ 36 | $\Sigma y^2 =$ 44 |

$$\bar{X} = \frac{\Sigma X}{N} = \frac{480}{8} = 60; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{552}{8} = 69$$

Karl Pearson's correlation coefficient,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \sqrt{\Sigma y^2}} \text{ where } \Sigma x = 0 \text{ and } \Sigma y = 0$$

$$= \frac{24}{\sqrt{36} \sqrt{44}}$$

$$= 0.6030$$

Example 2: Compute the coefficient of correlation between X - Advertisement Expenditure and Y - Sales.

| | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|
| X: | 10 | 12 | 18 | 8 | 13 | 20 | 22 | 15 | 5 | 17 |
| Y: | 88 | 90 | 94 | 86 | 87 | 92 | 96 | 94 | 88 | 85 |

(M.Com. Bharathidasan, N 92)

Solution:

Method I. Values of X and Y are assumed to be small and the following formula is attempted instead of the one used in the previous example.

| X | Y | XY | X ² | Y ² |
|--------------|--------------|---------------|----------------|----------------|
| | | 880 | 100 | 7744 |
| 10 | 88 | 1080 | 144 | 8100 |
| 12 | 90 | 1692 | 324 | 8836 |
| 18 | 94 | 688 | 64 | 7396 |
| 8 | 86 | 1131 | 169 | 7569 |
| 13 | 87 | 1840 | 400 | 8464 |
| 20 | 92 | 2112 | 484 | 9216 |
| 22 | 96 | 1410 | 225 | 8836 |
| 15 | 94 | 440 | 25 | 7744 |
| 5 | 88 | 1445 | 289 | 7225 |
| 17 | 85 | | | |
| $\Sigma X =$ | $\Sigma Y =$ | $\Sigma XY =$ | $\Sigma X^2 =$ | $\Sigma Y^2 =$ |
| 140 | 900 | 12718 | 2224 | 81130 |

Correlation coefficient,

$$\begin{aligned}
 r &= \frac{N \Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N \Sigma X^2 - (\Sigma X)^2} \sqrt{N \Sigma Y^2 - (\Sigma Y)^2}} \\
 &= \frac{10 \times 12718 - 140 \times 900}{\sqrt{10 \times 2224 - (140)^2} \sqrt{10 \times 81130 - (900)^2}} \\
 &= \frac{1180}{\sqrt{2640} \sqrt{1300}} \\
 &= 0.6370
 \end{aligned}$$

(or)

Method II. If the values of X and Y are large, the following formula can be used. This method is known as short-cut method or step deviation method or coded method. Any value can be assumed for a, b, c and d. If they are assumed as follows, the resulting values, will be smaller. $a = \bar{X}$ or some other convenient value in between the minimum and the maximum of X values. c is the common difference between X values. When there is no common difference, maximum possible value for c is to be identified such that u are not fractions. Similarly, from the values of Y, b and d are to be decided and $v = \frac{Y - b}{d}$ are to be calculated.

| X | Y | $u = \frac{X-a}{c}$ a=15; c=1 | $v = \frac{Y-b}{d}$ b=90; d=1 | uv | u^2 | v^2 |
|-------|-----|----------------------------------|----------------------------------|-----------------|------------------|------------------|
| 10 | 88 | -5 | -2 | 10 | 25 | 4 |
| 12 | 90 | -3 | 0 | 0 | 9 | 0 |
| 18 | 94 | 3 | 4 | 12 | 9 | 16 |
| 8 | 86 | -7 | -4 | 28 | 49 | 16 |
| 13 | 87 | -2 | -3 | 6 | 4 | 9 |
| 20 | 92 | 5 | 2 | 10 | 25 | 4 |
| 22 | 96 | 7 | 6 | 42 | 49 | 36 |
| 15 | 94 | 0 | 4 | 0 | 0 | 16 |
| 5 | 88 | -10 | -2 | 20 | 100 | 4 |
| 17 | 85 | 2 | -5 | -10 | 4 | 25 |
| Total | --- | $\sum u = -10$ | $\sum v = 0$ | $\sum uv = 118$ | $\sum u^2 = 274$ | $\sum v^2 = 130$ |

$$\begin{aligned}
 r &= \frac{N\sum uv - (\sum u)(\sum v)}{\sqrt{N\sum u^2 - (\sum u)^2} \sqrt{N\sum v^2 - (\sum v)^2}} \\
 &= \frac{10 \times 118 - (-10)(0)}{\sqrt{10 \times 274 - (-10)^2} \sqrt{10 \times 130 - (0)^2}} \\
 &= \frac{1180}{\sqrt{2640} \sqrt{1300}} \\
 &= 0.6370
 \end{aligned}$$

Note: r can be calculated by any formula in a problem. But, one formula will be easier than another.

Example 3: Calculate product moment correlation coefficient from the following bivariate frequency table.

| X \ Y | 1 | 3 | 5 |
|-------|---|---|---|
| -1 | 1 | 1 | 4 |
| 0 | 3 | 7 | 1 |
| 2 | 6 | 2 | 0 |

SPEARMAN'S RANK CORRELATION COEFFICIENT (ρ)

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2 - 1)} \right] \text{ when there is no tie. } d - \text{ difference between X and Y ranks.}$$

$$= 1 - \left[\frac{6 \left\{ \sum d^2 + \frac{m(m^2 - 1)}{12} \right\}}{N(N^2 - 1)} \right] \text{ when one value occurs } m \text{ times}$$

$$= 1 - \left[\frac{6 \left\{ \sum d^2 + \frac{m(m^2 - 1)}{12} + \frac{m(m^2 - 1)}{12} + \dots \right\}}{N(N^2 - 1)} \right] \text{ when more than one value is repeated}$$

It is calculated when ranks are given or when rank correlation coefficient is required. Rank correlation coefficient also lies between -1 and $+1$

Example 11: Rankings of 10 trainees at the beginning (X) and at the end (Y) of a certain course are given below:

| Trainees : | A | B | C | D | E | F | G | H | I | J |
|------------|---|---|---|---|---|---|---|----|---|----|
| X | 1 | 6 | 3 | 9 | 5 | 2 | 7 | 10 | 8 | 4 |
| Y | 6 | 8 | 3 | 7 | 2 | 1 | 5 | 9 | 4 | 10 |

Calculate Spearman's rank correlation coefficient

(I.C.W.A. Inter., J 95)

Solution :

| X | Y | d | d ² |
|----|----|----|----------------|
| 1 | 6 | -5 | 25 |
| 6 | 8 | -2 | 4 |
| 3 | 3 | 0 | 0 |
| 9 | 7 | 2 | 4 |
| 5 | 2 | 3 | 9 |
| 2 | 1 | 1 | 1 |
| 7 | 5 | 2 | 4 |
| 10 | 9 | 1 | 1 |
| 8 | 4 | 4 | 16 |
| 4 | 10 | -6 | 36 |

$$\rho = 1 - \left[\frac{6\sum d^2}{N(N^2 - 1)} \right]$$

$$= 1 - \left[\frac{6 \times 100}{10 \times 99} \right]$$

$$= 1 - 0.6061$$

$$= 0.3939$$

Total --- $\sum d = \sum d^2 =$

0 100

Example 12: X : 21 36 42 37 25
 Y : 47 40 37 42 43

For the data given above, calculate the rank correlation coefficient.

(B.Com. Bharathidasan, A 93)

Solution :

| X | Y | Rank | | d | d ² |
|-------|-----|------|-----|--------------|-----------------|
| | | X | Y | | |
| 21 | 47 | 5 | 1 | 4 | 16 |
| 36 | 40 | 3 | 4 | -1 | 1 |
| 42 | 37 | 1 | 5 | -4 | 16 |
| 37 | 42 | 2 | 3 | -1 | 1 |
| 25 | 43 | 4 | 2 | 2 | 4 |
| Total | --- | --- | --- | $\sum d = 0$ | $\sum d^2 = 38$ |

$$\rho = 1 - \left[\frac{6 \sum d^2}{N(N^2 - 1)} \right]$$

$$= 1 - \left[\frac{6 \times 38}{5(5^2 - 1)} \right]$$

$$= 1 - \left[\frac{6 \times 38}{5 \times 24} \right]$$

$$= 1 - 1.9$$

$$= -0.9$$

Note: For the maximum value of X, 42, rank is 1; for the next lower value 37, rank is 2; Similarly, for 47 of Y, rank is 1, 43 rank is 2,

2. Rank 1 may be assigned to the least value of X; rank 2 to the next higher value, . . . If so, the least value of Y is to be assigned rank 1, the next higher value rank 2 (Refer to Example 24).

Tied Ranks: When one or more values are repeated, the two aspects - ranks of the repeated values and change in the formula, are to be considered.

Each repeated value is to be considered separately. If a value has occurred m times, for each of them the average of the probable ranks which would have been assigned to them if they had differed slightly is assigned now. This does not affect the ranks of other values.

For each such repeated value, $\frac{m(m^2 - 1)}{12}$ is to be added with $\sum d^2$ once in the formula.

Example 13 : Find the rank correlation coefficient for the percentage of marks secured by a group of 8 students in Economics and Statistics.

Marks in Economics 50 60 65 70 75 40 70 80
 Marks in Statistics 80 71 60 75 90 82 70 50

Solution : Let, X - Marks in Economics,
 Y - Marks in Statistics.

| X | Y | Ranks | | d | d ² |
|-----------|----|-------|---|------|-------------------|
| | | X | Y | | |
| 50 | 80 | 7 | 3 | 4 | 16 |
| 60 | 71 | 6 | 5 | 1 | 1 |
| 65 | 60 | 5 | 7 | -2 | 4 |
| 70 | 75 | 3.5 | 4 | -0.5 | 0.25 |
| 75 | 90 | 2 | 1 | 1 | 1 |
| 40 | 82 | 8 | 2 | 6 | 36 |
| 70 | 70 | 3.5 | 6 | -2.5 | 6.25 |
| 80 | 50 | 1 | 8 | -7 | 49 |
| Total --- | | --- | | ∑d = | ∑d ² = |
| | | | | 0 | 113.5 |

$$\rho = 1 - \frac{6 \left\{ \sum d^2 + \frac{m(m^2 - 1)}{12} \right\}}{N(N^2 - 1)}$$

When m = 2, $\frac{m(m^2 - 1)}{12} = 0.5$

$$\therefore \rho = 1 - \frac{6\{113.5 + 0.5\}}{8(8^2 - 1)}$$

$$= 1 - \frac{6 \times 114}{8 \times 63}$$

$$= 1 - 1.3571 = -0.3571$$

Example 14: Marks obtained by 8 students in Accountancy (X) and Statistics (Y) are given below. Compute rank correlation.

| | | | | | | | | |
|---|----|----|----|----|----|----|----|----|
| X | 15 | 20 | 28 | 12 | 40 | 60 | 20 | 80 |
| Y | 40 | 30 | 50 | 30 | 20 | 10 | 30 | 60 |

(B.B.M. Bharathiar, A 95)

Solution:

| X | Y | Ranks | | d | d ² |
|-----------|----|-------|---|-----|-----------------|
| | | X | Y | | |
| 15 | 40 | 7 | 3 | 4 | 16 |
| 20 | 30 | 5.5 | 5 | 0.5 | 0.25 |
| 28 | 50 | 4 | 2 | 2 | 4 |
| 12 | 30 | 8 | 5 | 3 | 9 |
| 40 | 20 | 3 | 7 | -4 | 16 |
| 60 | 10 | 2 | 8 | -6 | 36 |
| 20 | 30 | 5.5 | 5 | 0.5 | 0.25 |
| 80 | 60 | 1 | 1 | 0 | 0 |
| Total --- | | --- | | ∑d | ∑d ² |
| | | | | = 0 | = 81.5 |

$$\rho =$$

$$1 - \frac{6 \left\{ \sum d^2 + \frac{m(m^2 - 1)}{12} + \frac{m(m^2 - 1)}{12} \right\}}{N(N^2 - 1)}$$

$$= 1 - \frac{6\{81.5 + 0.5 + 2\}}{8(8^2 - 1)}$$

$$= 1 - \frac{6 \times 84}{8 \times 63}$$

$$= 0$$

Note :

| Item | Freq. | Probable Ranks | Rank Assigned | $\frac{m(m^2-1)}{12}$ |
|--------|-------|----------------|-------------------------|-----------------------------|
| X = 20 | m = 2 | 5, 6 | $\frac{5+6}{2} = 5.5$ | $\frac{2(2^2-1)}{12} = 0.5$ |
| Y = 30 | m = 3 | 4, 5, 6 | $\frac{4+5+6}{3} = 5.0$ | $\frac{3(3^2-1)}{12} = 2.0$ |

Example 15: Ten competitors in a musical test were ranked by three judges A, B and C in the following order:

| | | | | | | | | | | |
|------------|---|---|---|----|---|----|---|----|---|---|
| Ranks by A | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9 | 7 | 8 |
| Ranks by B | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
| Ranks by C | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings in music.

Solution: When two judges have common likings they assign the same rank and consequently each $d=0$, $\sum d^2=0$ and $\rho = 1$. But no two judges have assigned the same ranks. For the pair of judges which has the nearest approach to common likings, ρ is nearest to +1.

| Ranks | | | | | | | | |
|-------|----|----|-------------------|-----------------------|-------------------|----------------------|-------------------|-----------------------|
| A | B | C | d_{AB} | d_{AB}^2 | d_{AC} | d_{AC}^2 | d_{BC} | d_{BC}^2 |
| 1 | 3 | 6 | -2 | 4 | -5 | 25 | -3 | 9 |
| 6 | 5 | 4 | 1 | 1 | 2 | 4 | 1 | 1 |
| 5 | 8 | 9 | -3 | 9 | -4 | 16 | -1 | 1 |
| 10 | 4 | 8 | 6 | 36 | 2 | 4 | -4 | 16 |
| 3 | 7 | 1 | -4 | 16 | 2 | 4 | 6 | 36 |
| 2 | 10 | 2 | -8 | 64 | 0 | 0 | 8 | 64 |
| 4 | 2 | 3 | 2 | 4 | 1 | 1 | -1 | 1 |
| 9 | 1 | 10 | 8 | 64 | -1 | 1 | -9 | 81 |
| 7 | 6 | 5 | 1 | 1 | 2 | 4 | 1 | 1 |
| 8 | 9 | 7 | -1 | 1 | 1 | 1 | 2 | 4 |
| Total | | | $\sum d_{AB} = 0$ | $\sum d_{AB}^2 = 200$ | $\sum d_{AC} = 0$ | $\sum d_{AC}^2 = 60$ | $\sum d_{BC} = 0$ | $\sum d_{BC}^2 = 214$ |

$$\rho_{AB} = 1 - \frac{6\sum d^2_{AB}}{N(N^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = -0.2121$$

$$\rho_{AC} = 1 - \frac{6\sum d^2_{AC}}{N(N^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 0.6364$$

$$\rho_{BC} = 1 - \frac{6\sum d^2_{BC}}{N(N^2 - 1)} = 1 - \frac{6 \times 214}{10 \times 99} = -0.2970$$

ρ_{AC} is nearest to +1 and so the pair A and C of judges has the nearest approach to common likings in music.

COEFFICIENT OF CORRELATION BY CONCURRENT DEVIATION METHOD (r_c)

$$r_c = \sqrt{\frac{2C - N}{N}} \quad \text{when } 2C - N > 0$$

$$= 0 \quad \text{when } 2C - N = 0$$

$$= -\sqrt{\frac{2C - N}{N}} \quad \text{when } 2C - N < 0$$

$$\therefore r_c = \pm \sqrt{\pm \left(\frac{2C - N}{N} \right)}$$

N denotes the number of entries and C denotes number of + signs (concurrent deviations) in D_{XY} column.

r_c also lies between -1 and +1.

If a value is greater than the preceding value, + sign is put. If it is less than the preceding one, - sign is marked. If it is equal to the preceding one, deviation is 0. D_X denotes such deviations among the values of the variable X and D_Y denotes those of Y . D_{XY} denotes the product of the entries under D_X and D_Y .

Example 16: Calculate the coefficient of correlation from the data given below by the method of concurrent deviations.

| | | | | | | |
|------------------|------|------|------|------|------|------|
| Year | 1959 | 1960 | 1961 | 1962 | 1963 | 1964 |
| Index of Imports | 85 | 82 | 89 | 95 | 104 | 108 |
| Index of Prices | 110 | 115 | 112 | 118 | 120 | 109 |
| Year | 1965 | 1966 | 1967 | 1968 | 1969 | |
| Index of Imports | 112 | 100 | 99 | 93 | 90 | |
| Index of Prices | 98 | 102 | 130 | 105 | 107 | |

SIMPLE LINEAR REGRESSION

13

Scatter diagram, explained in the previous chapter, helps to ascertain the nature of relationship between two variables. The relationship may be linear or non-linear. The discussion in this chapter is restricted to linear relationship. The method of least squares enables us to fit such an equation. The value of a related variable can be estimated corresponding to any value of a variable by using the equation.

While studying the hereditary characteristics in 1877, Sir Francis Galton used the word regression. The meaning of the word regression is *returning or going back*. Sir Francis Galton found that tall fathers had tall sons and short fathers had short sons. Further, the average height of the sons of the tall fathers was less than that of the tall fathers; the average height of the sons of the short fathers was more than that of the short fathers. The tendency of the average heights of different sections of sons was to move towards the average height of all the fathers. He preferred to call the tendency regression. The line which gives the average relationship between two variables is known as the regression line. The corresponding equation is the regression equation. The regression equation is also called *estimating equation*. The value of the dependent variable is estimated corresponding to any value of the independent variable by using the regression equation.

More than two interrelated variables are considered under multiple regression. Simple regression pertains to two variables. Linear relation between two variables is of interest under simple linear regression.

Uses

1. Regression analysis is used in Statistics and other disciplines. There are plenty of situations in which two or more variables are found to be interrelated. The tools to find the relation and to use it for prediction are provided by regression analysis.

2. The value of the dependent variable is estimated corresponding to any value of the independent variable using the appropriate regression equation.

3. In social research, the relation between variables may not be known; or, the relation may differ from place to place.

The relation can be established. Its validity can be ascertained by the test for goodness of fit.

4. In Economics and Business, there are many groups of interrelated variables. Regression analysis facilitates their study. For example, price, demand and supply are interrelated. The probable demand can be determined corresponding to any specified price and supply. Similarly, profit, revenue and expenditure on advertisement are interrelated. The probable amount of revenue can be predicted on the basis of any specified profit and expenditure on advertisement.

5. Regression analysis is of practical use in determining demand curve, supply curve, consumption function, etc. from market survey.

Correlation and Regression

Correlation coefficient indicates the direction of covariation and the closeness of the linear relation between two variables. If two variables are related, the mathematical equation of their relation is regression.

Regression equation gives the value of the dependent variable corresponding to any specified value of the independent variable.

The differences between correlation and regression are given in the following table:

| Correlation | Regression |
|---|---|
| 1. Correlation is the relationship between variables. It is expressed numerically. | 1. Regression means going back. The average relation between the variables is given as an equation. |
| 2. Between two variables, none is identified as independent or dependent variable. | One of the variables is independent variable and the other is dependent variable in any particular context. |
| 3. Correlation does not mean causation. One variable need not be the cause and the other, effect. | Independent variable may be 'the cause' and dependent variable, 'the effect'. |

4. There is spurious or nonsense correlation.

There is no such possibility. Regression is considered only when the variables are related.

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

Regression coefficients are independent of change of origin but are affected by change of scale.

6. Correlation coefficient is a number between - 1 and +1.

The two regression coefficients have the same sign, + or -. One of them can be greater than 1 numerically. But they can not be greater than 1 numerically simultaneously.

7. Correlation coefficient is not in any unit of measurement.

Each regression coefficient is in the unit of measurement of the dependent variable.

8. Correlation coefficient indicates the direction of covariation and the closeness of the linear relation between two variables.

Regression equations give the value of the dependent variable corresponding to any value of the independent variable.

9. The significance of the sample correlation coefficient can be tested. The limits between which the population correlation coefficient is expected to lie can be found.

Target can be reached. The value of the independent variable can be chosen so as to get the target value of the dependent variable. For example, a specific amount can be spent on advertisement to get the targeted revenue.

Two Regression Lines

For the pairs of values of X and Y, there are two regression lines.

When X is the independent variable and Y is the dependent variable, the line is called the regression line of Y on X. It is obtained by using the method of least squares as follows:

Let $Y = a + bX$ be the regression equation of Y on X .

Consider the observed pairs of values

$$(x_i, y_i) \quad i = 1, 2, 3, \dots$$

When $X = x_i$, $Y = a + bx_i$ from the above equation $a + bx_i$ is the estimated value of Y while the observed value is y_i .

$\therefore y_i - (a + bx_i) = y_i - a - bx_i$ is called error or residual or deviation. $D_1 = y_1 - a - bx_1$, $D_2 = y_2 - a - bx_2, \dots$ have been marked in the graph, Regression Line of Y on X . (Refer to the example 7). They are the vertical differences between the points $(x_1, y_1), (x_2, y_2), \dots$ and their foots on the line.

As explained in the chapter 'Method of Least Squares', the values of a and b for the given pairs of values of (x_i, y_i) $i = 1, 2, 3, \dots$ are determined such that the error sum of squares $E = D_1^2 + D_2^2 + D_3^2 + \dots$ is least. The conditions for the same are

$$\sum y = Na + b \sum x \text{ and}$$

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

and they are called normal equation. By solving these two equations, the values of a and b are found. They are, substituted in $Y = a + bX$ which is the required equation.

When Y is the independent variable and X is the dependent variable, the line is called the regression line of X on Y . It is obtained by using the method of least squares as follows:

Let $X = a' + b'Y$ be the regression equation of X on Y .

Consider the observed pairs of values

$$(x_i, y_i) \quad i = 1, 2, 3, \dots$$

When $Y = y_i$, $X = a' + b'y_i$ from the above equation. $a' + b'y_i$ is the estimated value of X while x_i is the observed value.

$\therefore x_i - (a' + b'y_i) = x_i - a' - b'y_i$ is called error or residual or deviation. $D_1 = x_1 - a' - b'y_1$, $D_2 = x_2 - a' - b'y_2, \dots$ have been marked in the graph, Regression Line of X on Y . (Refer to the Example 7). They are horizontal differences between the points $(x_1, y_1), (x_2, y_2), \dots$ and the corresponding points on the regression line of X on Y whose Y coordinates are y_1, y_2, \dots

Similar to the method of finding the values of a and b of the regression equation of Y on X , a' and b' are found by solving the normal equations

$$\sum x = Na' + b' \sum y \text{ and}$$

$$\sum xy = a' \sum y + b' \sum y^2$$

They are substituted in $X = a' + b' Y$ which is the required regression equation of X on Y .

For two variables, there is only one correlation coefficient but there are two regression lines. Why? It is a genuine question that arises in the minds of almost all people who study these aspects for the first time.

Correlation coefficient (r) between two variables is a number between -1 and $+1$. The sign indicates the direction of covariation of the variables. When the sign is negative, the two variables move in opposite directions. When the sign is positive, the two variables move in the same direction. The quantity indicates the extent of linear relation. When the pairs of values of the variables lie exactly on a line, $r = \pm 1$. When the pairs of values of the variables move away from a central line, r decreases numerically. Correlation does not distinguish between two variables as independent variable and dependent variable.

When two variables are related, the mathematical form of their relation is regression equation. Regression equation gives the value of the dependent variable corresponding to any specified value of the independent variable. When Y is the dependent variable, the method of least squares provides one equation. It is called the regression equation of Y on X . It is based on the vertical differences between the points (pairs of values) and the regression line. When X is the dependent variable, the method of least squares provides another equation. It is called the regression equation of X on Y . It is based on the horizontal differences between the points (pairs of values) and the regression line.

Generally there are two regression lines. But, when $r = -1$ or $+1$, they become one and the same line. When r decreases numerically they depart from one another. When r

decreases more and more, they depart further and further. When $r = 0$, they are perpendicular to each other. Regression line of Y on X is parallel to X axis and regression line of X on Y is parallel to Y axis.

Consider the variables price and demand. When price increases, demand is expected to decrease. (In this case, price is the independent variable and demand is the dependent variable). When demand decreases, price is also expected to decrease. (In this case, demand is the independent variable and price is the dependent variable). For the same two variables, there are two different relations. When price is the independent variable, there is one relation. When demand is the independent variable, there is another relation.

Methods of Forming The Regression Equations

Both the methods are based on the principle of least squares. They give the same equations.

1. Regression Equations on the basis of Normal Equations.

2. Regression Equations on the basis of \bar{X} , \bar{Y} , b_{XY} and b_{YX} .

Method 1. Regression Equations on the basis of Normal Equations.

Example 1: From the following data, obtain the two regression equations:

| | | | | | |
|---|---|----|----|---|---|
| X | 6 | 2 | 10 | 4 | 8 |
| Y | 9 | 11 | 5 | 8 | 7 |

(Use normal equations) (M.C.A. Bharathidasan, A 02)

Solution:

Steps: 1. A table is formed with the values of X and Y in the first two columns.

2. XY, X² and Y² are found and written in the next three columns.

3. Totals of the columns are found.

4. Normal equations for the regression equation of Y on X are considered. Values as per the table are substituted. The equations are then solved and the values of A and B are obtained.

By substituting those values in $Y = A + BX$ the required equation is obtained.

5. Normal equations for the regression equation of X on Y are considered next. Values as per the table are substituted. The equations are then solved and the values of A' and B' are obtained. By substituting those values in $X = A' + B'Y$ the required equation is obtained.

| X | Y | XY | X ² | Y ² | |
|--------------|----|--------------|----------------|-----------------|----------------|
| 6 | 9 | 54 | 36 | 81 | |
| 2 | 11 | 22 | 4 | 121 | |
| 10 | 5 | 50 | 100 | 25 | |
| 4 | 8 | 32 | 16 | 64 | |
| 8 | 7 | 56 | 64 | 49 | |
| $\Sigma x =$ | | $\Sigma y =$ | $\Sigma xy =$ | $\Sigma xy^2 =$ | $\Sigma y^2 =$ |
| 30 | | 40 | 214 | 220 | 340 |

Let the regression equation of Y on X be $Y = A + BX$

The normal equations are

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

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

By substituting the values from the table,

$$5A + 30B = 40 \text{ say (1)}$$

$$30A + 220B = 214 \dots (2)$$

$$(1) \times 6, \quad 30A + 180B = 240 \dots (3)$$

$$(2) - (3), \quad 40B = -26$$

$$\therefore B = \frac{-26}{40}$$

$$= -0.6500$$

$$\text{From (1), } 5A - 30 \times 0.6500 = 40$$

$$\therefore A = \frac{40 + 19.5}{5}$$

$$= 11.90$$

\therefore the regression equation of Y on X is

$$Y = 11.90 - 0.6500X$$

Let the regression equation of X on Y be $X = A' + B'Y$

The normal equations are

$$\Sigma X = NA' + B' \Sigma Y$$

$$\Sigma XY = A' \Sigma Y + B' \Sigma Y^2$$

By substituting the values from the table,

$$5A' + 40B' = 30 \text{ say (4)}$$

$$40A' + 340B' = 214 \dots (5)$$

$$40A' + 320B' = 240 \dots (6)$$

$$20B' = -26$$

$$B = \frac{-26}{20}$$

$$= -1.3000$$

$$\text{From (4), } 5A' + 40 \times (-1.30) = 30$$

$$A = \frac{30 + 52}{5}$$

$$= 16.40$$

\(\therefore\) the regression equation of X on Y is $X = 16.40 - 1.3000$

Method 2. Regression equations on the basis of \bar{X} , \bar{Y} , b_{XY} and b_{YX}

Regression equation of Y on X:

$$Y - \bar{Y} = b_{YX}(X - \bar{X})$$

b_{YX} is called the regression coefficient of Y on X.

$$b_{YX} = \frac{r\sigma_Y}{\sigma_X}$$

$$= \frac{\sum xy}{\sum x^2}$$

$$= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$

$$= \frac{d}{c} \times \frac{N\sum uv - (\sum u)(\sum v)}{N\sum u^2 - (\sum u)^2}$$

$$= \frac{N\sum f XY - (\sum f X)(\sum f Y)}{N\sum f X^2 - (\sum f X)^2}$$

$$= \frac{d}{c} \times \frac{N\sum f uv - (\sum f u)(\sum f v)}{N\sum f u^2 - (\sum f u)^2}$$

Regression equation of X on Y:

$$X - \bar{X} = b_{XY}(Y - \bar{Y})$$

b_{XY} is called the regression coefficient of X on Y.

$$b_{XY} = \frac{r\sigma_X}{\sigma_Y}$$

$$= \frac{\sum xy}{\sum y^2}$$

$$= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2}$$

$$= \frac{c}{d} \times \frac{N\sum uv - (\sum u)(\sum v)}{N\sum v^2 - (\sum v)^2}$$

$$= \frac{N\sum f XY - (\sum f X)(\sum f Y)}{N\sum f Y^2 - (\sum f Y)^2}$$

$$= \frac{c}{d} \times \frac{N\sum f uv - (\sum f u)(\sum f v)}{N\sum f v^2 - (\sum f v)^2}$$

\bar{X} , \bar{Y} , b_{XY} and b_{YX} are to be calculated by appropriate formulae and the regression equations are to be formed accordingly.

Example 2: You are given the following data:

| | | |
|---|----|------|
| Arithmetic mean | X | Y |
| Standard deviation | 36 | 85 |
| Correlation coefficient between X and Y | 11 | 8 |
| | | 0.66 |

- (a) Find the two regression equations.
- (b) Estimate the value of X when Y = 75.

(B.B.A. Bharathidasan, A 99; B.Com. Periyar, A 02)

Solution : Given: $\bar{X} = 36$; $\bar{Y} = 85$; $\sigma_X = 11$; $\sigma_Y = 8$; $r = 0.66$

$$\therefore b_{XY} = \frac{r\sigma_X}{\sigma_Y} = \frac{0.66 \times 11}{8} = 0.9075; b_{YX} = \frac{r\sigma_Y}{\sigma_X} = \frac{0.66 \times 8}{11} = 0.4800$$

(a) Regression equation of Y on X:

$$\begin{aligned}
 Y - \bar{Y} &= b_{YX} (X - \bar{X}) \\
 \text{i.e., } Y - 85 &= 0.4800 (X - 36) \\
 \text{i.e., } &= 0.4800 X - 17.28 \\
 \therefore Y &= 67.72 + 0.4800 X
 \end{aligned}$$

Regression equation of X on Y:

$$\begin{aligned}
 X - \bar{X} &= b_{XY} (Y - \bar{Y}) \\
 \text{i.e., } X - 36 &= 0.9075 (Y - 85) \\
 &= 0.9075 Y - 77.14 \\
 \therefore X &= 0.9075 Y - 41.14
 \end{aligned}$$

(b) When Y = 75, X = $0.9075 \times 75 - 41.14 = 26.92$

Note: If the value of Y is to be estimated, the regression equation of Y on X is to be used. For example, when X = 40, $Y = 67.72 + 0.4800 \times 40 = 86.92$

Example 3: From the following information on values of two variables X and Y find the two regression lines and the correlation coefficient:

$N = 10; \Sigma X = 20; \Sigma Y = 40; \Sigma X^2 = 240; \Sigma Y^2 = 410; \Sigma XY = 200$

Solution:

$$\bar{X} = \frac{\Sigma X}{N} = \frac{20}{10} = 2.00 \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{40}{10} = 4.00$$

$$b_{XY} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma Y^2 - (\Sigma Y)^2} \text{ as } \Sigma X \neq 0 \text{ and } \Sigma Y \neq 0$$

$$= \frac{10 \times 200 - 20 \times 40}{10 \times 410 - (40)^2}$$

$$= \frac{2000 - 800}{4100 - 1600}$$

$$= \frac{1200}{2500} = 0.4800$$

$$b_{YX} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2} \text{ as } \Sigma X \neq 0 \text{ and } \Sigma Y \neq 0$$

$$= \frac{1200}{10 \times 240 - (20)^2} \text{ from } b_{XY}$$

$$= \frac{1200}{2000} = 0.6000$$

Regression equation of Y on X :

$$Y - \bar{Y} = b_{YX} (X - \bar{X})$$

$$\text{i.e., } Y - 4 = 0.6000(X - 2)$$

$$= 0.6000 X - 1.20$$

$$\therefore Y = 2.80 + 0.6000X$$

Regression equation of X on Y :

$$X - \bar{X} = b_{XY} (Y - \bar{Y})$$

$$\text{i.e., } X - 2 = 0.4800 (Y - 4)$$

$$\text{i.e., } = 0.4800 Y - 1.92$$

$$\therefore X = 0.08 + 0.4800Y$$

The correlation coefficient,

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2} \sqrt{N\Sigma Y^2 - (\Sigma Y)^2}} \text{ as } \Sigma X \neq 0 \text{ and } \Sigma Y \neq 0$$

$$= \frac{1200}{\sqrt{2000} \sqrt{2500}} \text{ from } b_{XY} \text{ and } b_{YX}$$

$$= 0.5367 \quad (\text{or})$$

$$r = \pm \sqrt{b_{XY} \cdot b_{YX}} \text{ as mentioned later}$$

$$= + \sqrt{0.4800 \times 0.6000} \text{ as } b_{XY} \text{ and } b_{YX} \text{ are positive}$$

$$= 0.5367$$

Example 4: Calculate the two regression equations from the following data:

| | | | | | | |
|---|----|----|----|----|----|----|
| x | 10 | 12 | 13 | 12 | 16 | 15 |
| y | 40 | 38 | 43 | 45 | 37 | 43 |

Also estimate Y when X = 20.

(B.B.A. Bharathidasan, A O1)

Solution:

| X | Y | XY | X ² | Y ² |
|--------------|--------------|---------------|----------------|----------------|
| 10 | 40 | 400 | 100 | 1600 |
| 12 | 38 | 456 | 144 | 1444 |
| 13 | 43 | 559 | 169 | 1849 |
| 12 | 45 | 540 | 144 | 2025 |
| 16 | 37 | 592 | 256 | 1369 |
| 15 | 43 | 645 | 225 | 1849 |
| $\Sigma X =$ | $\Sigma Y =$ | $\Sigma XY =$ | $\Sigma X^2 =$ | $\Sigma Y^2 =$ |
| 78 | 246 | 3192 | 1038 | 10136 |

$$\bar{X} = \frac{\Sigma X}{N} = \frac{78}{6} = 13.00; \quad \bar{Y} = \frac{\Sigma Y}{N} = \frac{246}{6} = 41.00$$

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

$$= \frac{6 \times 3192 - 78 \times 246}{6 \times 10136 - (246)^2}$$

$$= \frac{19152 - 19188}{60816 - 60516} = \frac{-36}{300} = -0.1200$$

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

$$= \frac{-36}{6 \times 1038 - (78)^2} \text{ from } b_{XY}$$

$$= \frac{-36}{6228 - 6084} = \frac{-36}{144} = -0.2500$$

Regression equation of Y on X :

$$Y - \bar{Y} = b_{YX} (X - \bar{X})$$

$$\text{i.e., } Y - 41 = -0.2500 (X - 13)$$

$$= -0.2500 X + 3.25$$

$$Y = 44.25 - 0.25X$$

When $X = 20$, $Y = 44.25 - 0.25 \times 20 = 39.25$

Regression equation of X on Y :

$$X - \bar{X} = b_{XY} (Y - \bar{Y})$$

$$\text{i.e., } X - 13 = -0.1200 (Y - 41)$$

$$= -0.1200 Y + 4.92$$

$$\text{i.e., } X = 17.92 - 0.12Y$$

Example 5 : From the data given below, find:

- (a) the two regression equations
 (b) the coefficient of correlation between the marks in Mathematics and Statistics
 (c) the most likely marks in Statistics when the marks in Mathematics is 30.

Marks in

Mathematics (X): 25 28 35 32 31 36 29 38 34 32

Marks in

Statistics (Y) : 43 46 49 41 36 32 31 30 33 39

(B.Com. Periyar, A O2)

Solution:

| X | Y | $x = X - \bar{X}$ | $y = Y - \bar{Y}$ | xy | x^2 | y^2 |
|--------------|--------------|-------------------|-------------------|---------------|----------------|----------------|
| | | $\bar{X} = 32$ | $\bar{Y} = 38$ | | | |
| 25 | 43 | -7 | 5 | -35 | 49 | 25 |
| 28 | 46 | -4 | 8 | -32 | 16 | 64 |
| 35 | 49 | 3 | 11 | 33 | 9 | 121 |
| 32 | 41 | 0 | 3 | 0 | 0 | 9 |
| 31 | 36 | -1 | -2 | 2 | 1 | 4 |
| 36 | 32 | 4 | -6 | -24 | 16 | 36 |
| 29 | 31 | -3 | -7 | 21 | 9 | 49 |
| 38 | 30 | 6 | -8 | -48 | 36 | 64 |
| 34 | 33 | 2 | -5 | -10 | 4 | 25 |
| 32 | 39 | 0 | 1 | 0 | 0 | 1 |
| $\Sigma X =$ | $\Sigma Y =$ | $\Sigma x =$ | $\Sigma y =$ | $\Sigma xy =$ | $\Sigma x^2 =$ | $\Sigma y^2 =$ |
| 320 | 380 | 0 | 0 | -93 | 140 | 398 |