# Unit 3

# Correlation and Regression

**In this Unit we will learn**

1. Correlation
2. Types of Correlations
3. Methods of Studying Correlation
4. Scatter Diagram
5. Simple Graph
6. Karl Pearson's Coefficient of Correlation
7. Properties of Coefficient of Correlation
8. Rank Correlation
9. Regression
10. Types of Regression
11. Methods of Studying Regression
12. Lines of Regression
13. Regression Coefficients
14. Properties of Regression Coefficients
15. Properties of Lines of Regression (linear Regression)

## 1. Introduction

- Correlation and regression are statistical methods that are commonly used to compare two or more variables.
- For example, Comparison between income and expenditure, price and demand, etc.
- Correlation measures the association between two or more variables and quantitates the strength of their relationship. It evaluates only the existing data
- Regression means average relationship between two or more variables and this relationship is used to estimate the most likely values of one variable for specified values of the other variables.

## 2. Correlation

- Correlation is the relationship that exists between two or more variables. It is a statistical measure for finding out degree or strength of association between two or more variables.
- Two variables are said to be correlated if change in one variable affects a change in the other variable. Such a data connecting two variables is called bivariate data.

- Thus, correlation is a statistical analysis which measures and analyses the degree or extent to which two variables fluctuate with reference to each other.
- Some examples of such a relationship are as follows:

1. Relationship between heights and weights.
2. Relationship between age and strength of a body
3. Relationship between temperature and rain.

- The relationship between two variables may be linear or nonlinear
- If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable then correlation is said to be linear, otherwise it is called nonlinear.
- The linear correlation is measured by correlation coefficient or coefficient of correlation (r), which measures their degree of linear relationship between two quantitative variables.
- A significant advantage of the correlation coefficient is that it does not depend on the units of the variables and can therefore be used to compare any two variables regardless of their units.

## 3. Types of Correlations

Correlation is classified into four types:

I. Positive and negative correlations

2. Simple and multiple correlations

3. Partial and total correlations

4. Linear and nonlinear correlations

## 3.1 Positive and Negative Correlations

Depending on the variation in the variables, correlation may be positive or negative.

**1. Positive Correlation:**

If both the variables vary in the same direction, the correlation is said to be positive. In other words, if the value of one variable increases, the value of the other variable also increases, or, if value of one variable decreases, the value of the other variable decreases, e.g., the correlation between heights and weights of group of persons is a positive correlation

| Height(cm) | 150 | 157 | 163 | 170 | 178 |
|---|---|---|---|---|---|
| Weight(kg) | 58 | 62 | 68 | 73 | 80 |

**2. Negative Correlation:**

If both the variables vary in the opposite direction, correlation is said to be negative. In other words, if the value of one variable increases, the value of the other variable decreases, or, if the value of one variable decreases the value of the other variable increases, e.g. the correlation between the price and demand of a commodity is a negative correlation.

| Price ($per unit) | 10 | 8 | 6 | 5 | 4 |
|---|---|---|---|---|---|
| Demand (units) | 100 | 200 | 300 | 400 | 500 |

## 3.2 Simple and Multiple Correlations

Depending upon the study of the number of variables, correlation may be simple or multiple.

**1. Simple Correlation**

When only two variables are studied, the relationship is described as simple correlation. e.g., the quantity of money and price level, demand and price, etc.

**2. Multiple Correlation**

When more than two variables are studied, the relationship is described as multiple correlation, e.g., relationship of price, demand, and supply of a commodity,

## 3.3 Partial and Total Correlations

Multiple correlation may be either partial or total.

**1. Partial Correlation**

When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

**2. Total Correlation**

When more than two variables are studied without excluding any variables, the relationship is termed as total correlation.

## 3.4 Linear and Nonlinear Correlations

Depending upon the ratio of change between two variables, the correlation may be linear or nonlinear.

**1. Linear Correlation**

If the ratio of change between two variables is constant, the correlation is said to be linear. If such variables are plotted on a graph paper, a straight line is obtained, e.g.,

| | | | | | |
|---|---|---|---|---|---|
| Milk(litter) | 5 | 10 | 15 | 20 | 25 |
| Paneer (kg) | 2 | 4 | 6 | 8 | 10 |

**2. Nonlinear Correlation**

If the ratio of change between two variables is not constant, the correlation is said to nonlinear. The graph of a nonlinear or curvilinear relationship will be a curve, e.g.,

| | | | | | |
|---|---|---|---|---|---|
| Expenses (in lacs) | 3 | 6 | 9 | 12 | 15 |
| Sales (in lacs) | 8 | 12 | 15 | 15 | 16 |

## 4. Measures of Correlation or Method of Studying Correlation

Correlation between two variables $X$ and $Y$ can be calculated by any one of the following methods.

1. Two-way frequency table
2. Scatter diagram
3. Karl Pearson's coefficient of correlation
4. Spearman's rank correlation method.
5. Concurrent deviation method.

Here we will take the idea of methods (2), (3) and (4)

## 4.1 Scatter Diagram

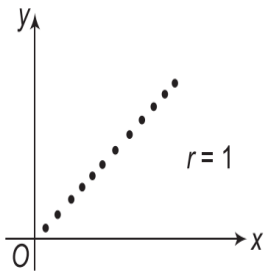
- The scatter diagram is a diagrammatic representation of bivariate data to find the correlation between two variables. In this case the first essential step for calculating correlation coefficient is to plot the observation in a scatter gram or scatter plot to visually evaluate the data for potential relationship or the presence of outlying values.

- The given statistical data is plotted as point on a rectangular Cartesian coordinate system taking one independent variable (say X) along horizontal axis while another dependent variable (say Y) along the vertical axis. Then we have the following interpretation regarding their nature and strength of correlation.

There are various correlationships between two variables represented by the following scatter diagrams.

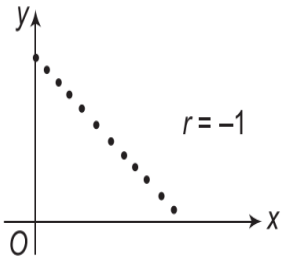
## 1. Perfect Positive Correlation

If all the plotted points lie on a straight line rising from the lower left-hand corner to the upper right-hand corner, the correlation is said to be perfectly positive.



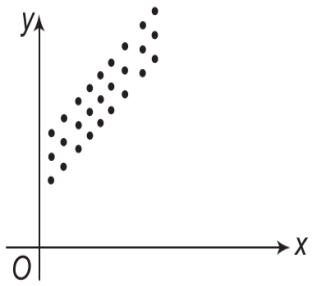

## 2. Perfect Negative Correlation

If all the plotted points lie on a straight line falling from the upper-left hand corner to the lower right-hand corner, the correlation is said to be perfectly negative.



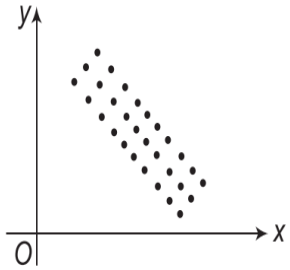

## 3. High Degree of Positive Correlation

If all the plotted points lie in the narrow strip, rising from the lower left-hand corner to the upper right-hand corner, it indicates a high degree of positive correlation.



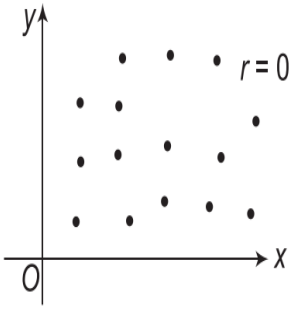

### 4. High Degree of Negative Correlation

If all the plotted points lie in a narrow strip, falling from the upper left-hand corner to the lower right-hand corner, it indicates the existence of a high degree of negative correlation.




### 5. No Correlation

If all the plotted points lie on a straight line parallel to the x-axis or y-axis or in a haphazard manner, it indicates the absence of any relationship between the variables.




## 4.2 Merits of a Scatter Diagram

1. It is simple and nonmathematical method to find out the correlation between the variables.

2. It gives an indication of the degree of linear correlation between the variables.

3. It is easy to understand.

4. It is not influenced by the size of extreme items.

## 4.3 Simple Graph

A simple graph is a diagrammatic representation of bivariate data to find the correlation between two variables. The values of the two variables are plotted on a graph paper. Two curves are obtained, one for the variable x and the other for the variable y. If both the curves move in the same direction, the correlation is said to be positive. If both the curves move in the opposite direction, the correlation is said to be negative. This method is used in the case of a time series. It does not reveal the extent to which the variables are related.

Mathematical methods are

(a) Karl Pearson's coefficient of correlation

(b) Spearman's rank coefficient correlation

## 4.4 Karl Pearson's Coefficient of Correlation (Covariance Method)

Before we introduce this method, let us define the concept of covariance.

### ➢ Covariance

The coefficient of correlation is the measure of correlation between two random variables X and Y, and is denoted by r.

Let $(x_1, y_1), (x_2,\ y_2), \dots, (x_n, y_n)$ be n-pair of observations on two variables $X$ and $Y$, then the covariance of $X$ and $Y$ is denoted and defined as

$$cov\ (X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

Where $\bar{x}$ and $\bar{y}$ are arithmetic means of $X$ and $Y$ series, respectively; that is

$$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}, \bar{y} = \frac{y_1+y_2+\cdots+y_n}{n}$$

Covariance indicates the join variations between the two variables.

## ➢ Karl Person's Coefficient of Correlation

The coefficient of correlation is the measure of correlation between two random variables $X$ and $Y$, and is denoted by $r$.

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

Where $cov\ (X,Y)$ is the covariance if variables $X$ and $Y$.

$\sigma_x$ is the standard deviation of variable $X$.

$\sigma_y$ is the standard deviation of variable $Y$.

This expression is known as Karl Pearson's coefficient of correlation or Karl Pearson's product-moment coefficient of correlation

$$cov(X,Y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

$$\therefore r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

The above expression can be further modified.

Expanding the terms,

$$r = \frac{\sum(xy-x\bar{y}-\bar{x}y+\bar{x}\bar{y}}{\sqrt{\sum(x^2-2x\bar{x}+\bar{x}^2)}\sqrt{\sum(y^2-2y\bar{y}+\bar{y}^2)}}$$

$$= \frac{\sum xy-\bar{y}\sum x-\bar{x}\sum y+\bar{x}\bar{y}\sum 1}{\sqrt{\sum x^2-2\bar{x}\sum x+\bar{x}^2\sum 1}\sqrt{\sum y^2-2\bar{y}\sum y+\bar{y}^2\sum 1}}$$

$$= \frac{\sum xy-\frac{\sum y}{n}\sum x-\frac{\sum x}{n}\sum y+\frac{\sum x\sum y}{n\ \ n}\sum 1}{\sqrt{\sum x^2-2\frac{\sum x}{n}\sum x+\left(\frac{\sum x}{n}\right)^2 n}\sqrt{\sum y^2-2\frac{\sum y}{n}\sum y+\left(\frac{\sum y}{n}\right)^2 n}}$$

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

## 4.5 Properties of Coefficient of Correlation

**1. The coefficient of correlation lies between -1 and 1, i.e., $-1 \leq r \leq 1$.**

**Proof:**

Let $\bar{x}$ and $\bar{y}$ be the mean of $x$ and $y$ series and $\sigma_x$ and $\sigma_y$ be their respective standard deviations.

Let $\sum\left(\frac{x-\bar{x}}{\sigma_x} \pm \frac{y-\bar{y}}{\sigma_y}\right)^2 \geq 0$ [$\because$ sum of squares of real quantities cannot be negative]

$$\frac{\sum(x-\bar{x})^2}{\sigma_x^2} + \frac{\sum(y-\bar{y})^2}{\sigma_y^2} \pm \frac{2\sum(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} \geq 0$$

$$\Rightarrow n + n \pm 2nr \geq 0$$

$$\Rightarrow 2n \pm 2nr \geq 0$$

$$\Rightarrow 2n(1 \pm r) \geq 0$$

$$\Rightarrow 1 \pm r \geq 0$$

$$\Rightarrow 1 + r \geq 0 \quad or \quad 1 - r \geq 0$$

$$\Rightarrow r \geq -1 \quad or \quad r \leq 1$$

Hence, the coefficient of correlation lies between -1 and 1. i.e., $-1 \leq r \leq 1$.

**2. Correlation coefficient is independent of change of origin and change of scale.**

**Proof:**

Let $d_x = \frac{x-a}{h}, \; d_y = \frac{y-b}{k}$

$$\therefore x = a + hd_x, \; y = b + kd_y$$

Where $a, b, h(>0)\; and\; k(>0)$ are constants.

$$x = a + hd_x \Rightarrow \bar{x} = a + h\overline{d_x} \Rightarrow x - \bar{x} = h(d_x - \overline{d_x})$$

$$y = b + kd_y \Rightarrow \bar{y} = b + h\overline{d_y} \Rightarrow y - \bar{y} = k(d_y - \overline{d_y})$$

$$r_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$= \frac{\sum h(d_x-\overline{d_x})k(d_y-\overline{d_y})}{\sqrt{\sum h^2(d_x-\overline{d_x})^2}\sqrt{\sum k^2(d_y-\overline{d_y})^2}}$$

$$= \frac{\sum(d_x-\overline{d_x})(d_y-\overline{d_y})}{\sqrt{\sum(d_x-\overline{d_x})^2}\sqrt{\sum(d_y-\overline{d_y})^2}}$$

$$= r_{d_xd_y}$$

Hence, the correlation coefficient is independent of change of origin and change of scale.

**Note:** Since correlation coefficient is independent of change of origin and change of scale,

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

**3. Two independent variables are uncorrelated.**

**Proof:**

If random variables $X$ and $Y$ are independent,

$$\sum(x - \bar{x})(y - \bar{y}) = 0 \text{ or } cov(X, Y) = 0$$

$$\therefore r = 0$$

Thus, if $X$ and $Y$ are independent variables, they are uncorrelated.

**Note:**

The converse of the above property is not true, i.e., two uncorrelated variables may not be independent

## 5 Examples

**Example-1: Calculate the correlation coefficient between $x$ and $y$ using the following data:**

| **X** | **2** | **4** | **5** | **6** | **8** | **11** |
|---|---|---|---|---|---|---|
| **y** | **18** | **12** | **10** | **8** | **7** | **5** |

**Solution:**

$n = 6$

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 2 | 18 | 4 | 324 | 36 |
| 4 | 12 | 16 | 144 | 48 |
| 5 | 10 | 25 | 100 | 50 |
| 6 | 8 | 36 | 64 | 48 |
| 8 | 7 | 64 | 49 | 56 |
| 11 | 5 | 121 | 25 | 55 |
| $\sum x = 36$ | $\sum y = 60$ | $\sum x^2 = 266$ | $\sum y^2 = 706$ | $\sum xy = 293$ |

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

$$= \frac{293 - \frac{(36)(60)}{6}}{\sqrt{266 - \frac{(36)^2}{6}}\sqrt{706 - \frac{(60)^2}{6}}}$$

$$= -0.9203$$

**Note $\sum x, \sum y, \sum x^2, \sum y^2, \sum xy$ can be directly obtained with the help of scientific calculator.**

**<u>Example-2</u>: Calculate the correlation coefficient between $x$ and $y$ using the following data:**

| **X** | **5** | **9** | **13** | **17** | **21** |
|---|---|---|---|---|---|
| **y** | **12** | **20** | **25** | **33** | **35** |

**Solution:**

$n = 5$

$\bar{x} = \frac{\sum x}{n} = \frac{65}{5} = 13$

$\bar{y} = \frac{\sum y}{n} = \frac{125}{5} = 25$

| $x$ | $y$ | $x - \bar{x}$ | $y - \bar{y}$ | $(x - \bar{x})^2$ | $(y - \bar{y})^2$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|---|---|
| 5 | 12 | -8 | -13 | 64 | 169 | 104 |
| 9 | 20 | -4 | -5 | 16 | 25 | 20 |
| 13 | 25 | 0 | 0 | 0 | 0 | 0 |
| 17 | 33 | 4 | 8 | 16 | 64 | 32 |
| 21 | 35 | 8 | 10 | 64 | 100 | 80 |
| $\sum x = 65$ | $\sum y = 125$ | $\sum (x - \bar{x}) = 0$ | $\sum (y - \bar{y}) = 0$ | $\sum (x - \bar{x})^2 = 160$ | $\sum (y - \bar{y})^2 = 358$ | $\sum (x - \bar{x})(y - \bar{y}) = 236$ |

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2}\sqrt{\sum (y - \bar{y})^2}}$$

$$= \frac{236}{\sqrt{160}\sqrt{358}}$$

$$= 0.986$$

**<u>Example-3</u>: Calculate the coefficient of correlation for the following pairs of $x$ and $y$.**

| **X** | **17** | **19** | **21** | **26** | **20** | **28** | **26** | **27** |
|---|---|---|---|---|---|---|---|---|
| **y** | **23** | **27** | **25** | **26** | **27** | **25** | **30** | **33** |

**Solution:**

Let $a = 23$ and $b = 27$ be the assumed means of $x$ and $y$ series respectively.

$d_x = x - a = x - 23$

$d_y = y - b = y - 27$

$n = 8$

| $x$ | $y$ | $d_x$ | $d_y$ | $d_x^2$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 17 | 23 | -6 | -4 | 36 | 16 | 24 |
| 19 | 27 | -4 | 0 | 16 | 0 | 0 |
| 21 | 25 | -2 | -2 | 4 | 4 | 4 |
| 26 | 26 | 3 | -1 | 9 | 1 | -3 |
| 20 | 27 | -3 | 0 | 9 | 0 | 0 |
| 28 | 25 | 5 | -2 | 25 | 4 | -10 |
| 26 | 30 | 3 | 3 | 9 | 9 | 9 |
| 27 | 33 | 4 | 6 | 16 | 36 | 24 |
| | | $\sum d_x = 0$ | $\sum d_y = 0$ | $\sum d_x^2 = 160$ | $\sum d_y^2 = 358$ | $\sum d_x d_y = 236$ |

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{48-0}{\sqrt{124-0}\sqrt{70-0}}$$

$$= 0.515$$

**<u>Example-4</u>: Calculate the Karl Pearson's coefficient of correlation between the ages of cars and annual maintenance costs.**

| **Ages of cars(year)** | **2** | **4** | **6** | **7** | **8** | **10** | **12** |
|---|---|---|---|---|---|---|---|
| **Annual maintenance cost (Rupee)** | **1600** | **1500** | **1800** | **1900** | **1700** | **2100** | **2000** |

**Solution:**

Let the ages of cars in years be denoted by $x$ and annual maintenance costs in rupees be denoted by $y$.

Let $a = 7$ and $b = 1800$ be assumed means of $x$ and $y$ series respectively.

Let $h = 1, k = 100$

$$d_x = \frac{x-a}{h} = \frac{x-7}{1} = x - 7$$

$$d_y = \frac{y-b}{k} = \frac{y-1800}{100}$$

$$n = 7$$

| $x$ | $y$ | $d_x$ | $d_y$ | $d_x^2$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 2 | 1600 | -5 | -2 | 25 | 4 | 10 |
| 4 | 1500 | -3 | 3 | 9 | 9 | 9 |
| 6 | 1800 | -1 | 0 | 1 | 0 | 0 |
| 7 | 1900 | 0 | 1 | 0 | 1 | 0 |
| 8 | 1700 | 1 | -1 | 1 | 1 | -1 |
| 10 | 2100 | 3 | 3 | 9 | 9 | 9 |
| 12 | 2000 | 5 | 2 | 25 | 4 | 10 |
| | | $\sum d_x = 0$ | $\sum d_y = 0$ | $\sum d_x^2 = 70$ | $\sum d_y^2 = 28$ | $\sum d_x d_y = 37$ |

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}} \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{37-0}{\sqrt{70-0}\sqrt{28-0}}$$

$$= 0.836$$

**Example-5: The coefficient of correlation between two variables $X$ and $Y$ is $0.48$. The covariance is 36. The variance of $X$ is 16. Find the Standard deviation of $Y$.**

**Solution**:

$r = 0.48, \quad cov(X,Y) = 36, \quad \sigma_x^2 = 16$

$\therefore \sigma_X = 4$

$$r = \frac{cov\ (X,Y)}{\sigma_X \sigma_Y}$$

$$0.48 = \frac{36}{4\sigma_y}$$

$\therefore \sigma_y = 18.75$

**Example-6: From the following information, calculate the value of $n$.**

$\sum x = 4, \sum y = 4, \ \sum x^2 = 44, \ \sum y^2 = 44, \ \sum xy = -40, \ r = -1$

**Solution:**

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

$$-1 = \frac{-40 - \frac{(4)(4)}{n}}{\sqrt{44 - \frac{(4)^2}{n}}\sqrt{44 - \frac{(4)^2}{n}}}$$

$\therefore n = 8$

**Example-7: Calculate the correlation coefficient between $x$ and $y$ from the following data:**

$$n = 10, \sum x = 140, \sum y = 150, \sum (x-10)^2 = 180$$

$$\sum (y-15)^2 = 215, \sum (x-10)(y-15) = 60$$

**Solution:**

$\sum d_x^2 = \sum (x-10)^2 = 180$

$\sum d_y^2 = \sum (y-15)^2 = 215$

$\sum d_x d_y = \sum (x-10)(y-15) = 60$

$a = 10$

$b = 15$

$n = 10$

$$\bar{x} = \frac{\sum x}{n} = \frac{140}{10} = 14$$

$$\bar{y} = \frac{\sum y}{n} = \frac{150}{10} = 15$$

$$\bar{x} = a + \frac{\sum d_x}{n} \Rightarrow 14 = 10 + \frac{\sum d_x}{10} \Rightarrow \sum d_x = 40$$

$$\bar{y} = b + \frac{\sum d_y}{n} \Rightarrow 15 = 15 + \frac{\sum d_y}{10} \Rightarrow \sum d_y = 0$$

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}} \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$= \frac{60 - \frac{(40)(0)}{10}}{\sqrt{180 - \frac{(40)^2}{10}} \sqrt{215 - \frac{(0)^2}{10}}}$$

$$= 0.915$$

**<u>Example-8</u>: A computer operator while calculating the coefficient between two variates $x$ and $y$ for 25 pairs of observation obtained the following constants:**

$$\boldsymbol{n = 25, \sum x = 125, \sum x^2 = 650, \sum y = 100, \sum y^2 = 460, \sum xy = 508}$$

**It was later discovered at the time of checking that he had copied down two pairs as $(6, 14)$ and $(8, 6)$ while the correct pair were $(8, 12)$ and $(6, 8)$. Obtain the correct value of the correlation coefficient.**

**Solution:**

$n = 25$

Corrected $\sum x$ = Incorrected $\sum x$ −(sum of incorrected $x$) + (sum of corrected $x$)

$$= 125 - (6 + 8) + (8 + 6)$$

$$= 125$$

Similarly,

Corrected $\sum y = 100 - (14 + 6) + (12 + 8) = 100$

Corrected $\sum x^2 = 650 - (6^2 + 8^2) + (8^2 + 6^2) = 650$

Corrected $\sum y^2 = 460 - (14^2 + 6^2) + (12^2 + 8^2) = 436$

Corrected $\sum xy = 508 - (84 + 48) + (96 + 48) = 520$

Correct value of correlation coefficient

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}\sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}} = \frac{520 - \frac{(125)(100)}{25}}{\sqrt{650 - \frac{(125)^2}{25}}\sqrt{436 - \frac{(100)^2}{25}}} = 0.67$$

## 6. Rank Correlation

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

### 6.1 Spearman's Rank Correlation Coefficient

Let $x, y$ be the rank of the $i^{th}$ individuals in two characteristics $A$ and $B$ respectively where $i = 1,2, \dots, n$. Assuming that no two individuals have the same rank either for $x$ or $y$, each of the variables $x$ and $y$ take the values $1,2, \dots, n$.

$$\bar{x} = \bar{y} = \frac{1 + 2 + 3 + \cdots + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

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

$$= \sum x^2 - 2\bar{x}\sum x + \bar{x}^2 \sum 1$$

$$= \sum x^2 - 2n\bar{x} + n\bar{x}^2 \qquad [\because \sum x = n\bar{x}\ ]$$

$$= \sum x^2 - n\bar{x}^2$$

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

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

$$= \frac{1}{12}(n^3 - n)$$

Similarly, $\sum (y - \bar{y})^2 = \frac{1}{12}(n^3 - n)$

If $d$ denoted the difference between the rank of the $i^{th}$ individuals in the two variables,

$$d = x - y = (x - \bar{x}) - (y - \bar{y}) \qquad [\because\ \bar{x} = \bar{y}\ ]$$

Squaring and summing over $i$ from 1 to $n$,

$$\sum d^2 = \sum [(x - \bar{x}) - (y - \bar{y})]^2$$

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

$$\sum (x - \bar{x}) - (y - \bar{y}) = \frac{1}{2}[\sum (x - \bar{x})^2 + \sum (y - \bar{y})^2 - \sum d^2] = \frac{1}{12}(n^3 - n) - \frac{1}{2}\sum d^2$$

Hence, the coefficient of correlation between these variables is

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

$$= \frac{\frac{1}{12}(n^3-n)-\frac{1}{2}\sum d^2}{\frac{1}{12}(n^3-n)}$$

$$= 1 - \frac{6\sum d^2}{n^3-n}$$

$$= 1 - \frac{6\sum d^2}{n(n^2-1)}$$

This is called Spearman's rank correlation coefficient and is denoted by $\rho$.

**<u>Example-8</u>: Ten participants in a contest are ranked by two judges as follows**

| $x$ | 1 | 3 | 7 | 5 | 4 | 6 | 2 | 10 | 9 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 3 | 1 | 4 | 5 | 6 | 9 | 7 | 8 | 10 | 2 |

**Calculate the rank correlation coefficient.**

**Solution:**

$n = 10$

| Rank by first Judge $x$ | Rank by second judge $y$ | $d = x - y$ | $d^2$ |
|---|---|---|---|
| 1 | 3 | -2 | 4 |
| 3 | 1 | 2 | 4 |
| 7 | 4 | 3 | 9 |
| 5 | 5 | 0 | 0 |
| 4 | 6 | -2 | 4 |
| 6 | 9 | -3 | 9 |
| 2 | 7 | -5 | 25 |
| 10 | 8 | 2 | 4 |
| 9 | 10 | -1 | 1 |
| 8 | 2 | 6 | 36 |
| | | $\sum d = 0$ | $\sum d^2 = 96$ |

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(96)}{10[(10)^2-1]} = 0.418$$

**<u>Example-9</u>: Ten competitors in a musical test were ranked by the three judges $A, B$ and $C$ in the following order:**

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

**Using the rank correlation method, find which pair of judges has the nearest approach to common liking in music.**

**Solution:**

$n = 10$

| Rank by A $x$ | Rank by B $y$ | Rank by C $z$ | $d_1 = x - y$ | $d_2 = y - z$ | $d_3 = z - x$ | $d_1^2$ | $d_2^2$ | $d_3^2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 6 | -2 | -3 | 5 | 4 | 9 | 25 |
| 6 | 5 | 4 | 1 | 1 | -2 | 1 | 1 | 4 |
| 5 | 8 | 9 | -3 | -1 | 4 | 9 | 1 | 16 |
| 10 | 4 | 8 | 6 | -4 | -2 | 36 | 16 | 4 |
| 3 | 7 | 1 | -4 | 6 | -2 | 16 | 36 | 4 |
| 2 | 10 | 2 | -8 | 8 | 0 | 64 | 64 | 0 |
| 4 | 2 | 3 | 2 | -1 | -1 | 4 | 1 | 1 |
| 6 | 1 | 10 | 8 | -9 | 1 | 64 | 81 | 1 |
| 7 | 6 | 5 | 1 | 1 | -2 | 1 | 1 | 4 |
| 8 | 9 | 7 | -1 | 2 | -1 | 1 | 4 | 4 |
| | | | $\sum d_1 = 0$ | $\sum d_2 = 0$ | $\sum d_3 = 0$ | $\sum d_1^2 = 200$ | $\sum d_2^2 = 214$ | $\sum d_3^2 = 60$ |

$$r(x,y) = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(200)}{10[(10)^2-1]}$$

$$= -0.21$$

$$r(y,z) = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(214)}{10[(10)^2-1]}$$

$$= -0.296$$

$$r(z,x) = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(60)}{10[(10)^2-1]}$$

$$= 0.64$$

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

## 6.2 Tied Rank

If there is a tie between two or more individuals rank, the rank is divided among equal individuals, e.g., if two items have fourth rank, the 4th and 5th rank is divided between them equally and is given as $\frac{4+5}{2} = 4.5^{th}$ rank to each of them. If three items have the same 4th rank, each of them is given $\frac{4+5+6}{3} = 5^{th}$ rank. As a result of this, the following adjustment or correlation is made in the rank correlation formula. If $m$ is the number of items having equal ranks then the factor $\frac{1}{12}(m^3 - m)$

Is added to $\sum d^2$. If there are more than one cases of this type, this factor is added corresponding to each case.

$$r = \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{n(n^2-1)}$$

**Example-10: Obtain the rank correlation coefficient from the following data:**

| $x$ | 10 | 12 | 18 | 18 | 15 | 40 |
|---|---|---|---|---|---|---|
| $y$ | 12 | 18 | 25 | 25 | 50 | 25 |

**Solution:**

Here, $n = 6$

| $x$ | $y$ | Rank $x$ | Rank $y$ | $d = x - y$ | $d^2$ |
|---|---|---|---|---|---|
| 10 | 12 | 1 | 1 | 0 | 0 |
| 12 | 18 | 2 | 2 | 0 | 0 |
| 18 | 25 | 4.5 | 4 | 0.5 | 0.25 |
| 18 | 25 | 4.5 | 4 | 0.5 | 0.25 |
| 15 | 50 | 3 | 6 | -3 | 9 |
| 40 | 25 | 6 | 4 | 2 | 4 |
| | | | | | $\sum d^2 = 13.5$ |

There are two items in the x series having equal values at the rank 4. Each is given the rank 4.5. Similarly, there are three items in the y series at the rank 3. Each of them is given the rank 4.

$$m_1 = 2, m_2 = 3$$

$$r = \frac{6\left[\sum d^2 + \frac{1}{12}(m_1^3 - m_1) + \frac{1}{12}(m_2^3 - m_2) + \cdots\right]}{n(n^2 - 1)}$$

$$= \frac{6\left[13.50 + \frac{1}{12}(8 - 2) + \frac{1}{12}(27 - 3) + \cdots\right]}{6(6^2 - 1)}$$

$$= 0.5429$$

## 7. Regression

Regression is defined as a method of estimating the value of one variable when that of the other is known and the variables are correlated. Regression analysis is used to predict or estimate one variable in terms of the other variable. It is a highly valuable tool for prediction purpose in economics and business. It is useful in statistical estimation of demand curves, supply curves, production function, cost function, consumption function, etc.

## 8. Types of Regression

Regression is classified into two types:

1. Simple and multiple regressions
2. Linear and nonlinear regressions

## 8.1 Simple and Multiple Regressions

Depending upon the study of the number of variables., regression may be simple or multiple.

**1. Simple Regression**

The regression analysis for studying only two variables at a time is known as simple regression.

**2. Multiple Regression**

The regression analysis for studying more than two variables at a time is known as multiple regression.

## 8.2 Linear and Nonlinear Regressions

Depending upon the regression curve, regression may be linear or nonlinear.

**1. Linear Regression**

If the regression curve is a straight line, the regression is said to be linear.

**2. Nonlinear Regression**

If the regression curve is not a straight line i.e., not a first-degree equation in the variables $x$ and $y$, the regression is said to be nonlinear or curvilinear, In this case, the regression equation will have a functional relation between the variables $x$ and $y$ involving terms in $x$ and $y$ of the degree higher than one i.e., involving terms of the type $x^2, y^2, x^3, y^3, xy$ etc.

## 9. Methods of Studying Regression

There are two methods of studying correlation:

(i) Method of scatter diagram

(ii) Method of least squares

## 9.1 Method of Scatter Diagram

It is the simplest method of obtaining the lines of regression. The data are plotted on a graph paper by taking the independent variable on the x-axis and the dependent variable on the y-axis. Each of these points are generally scattered in a narrow strip. If the correlation is perfect, i.e., if $r$ is equal to one, positive, or negative, the points will lie on a line which is the line of regression.

## 9.2 Method of Least Squares

This is a mathematical method which gives an objective treatment to find a line of regression. It is used for obtaining the equation of a curve which fits best to a given set of observations. It is based on the assumption that the sum of squares of differences between the estimated values and the actual observed values of the observations is minimum.

## 10. Lines of Regression

If the variables, which are highly correlated, are plotted on a graph then the points lie in a narrow strip. If all the points in the scatter diagram cluster around a straight line, the line is called the line of regression. The line of regression is the line of best fit and is obtained by the principle of least squares.

## 10.1 Line of Regression of y on x

It is the line which gives the best estimate for the values of $y$ for any given values of $x$. The regression equation of $y$ on $x$ is given by

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

It is also written as $y = a + bx$

## 10.2 Line of Regression of x on y

It is the line which gives the best estimate for the values of $x$ for any given values of $y$. The regression equation for $x$ on $y$ is given by

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

It is also written as $x = a + by$

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

## 11. Regression Coefficients

The slope $b$ of the line of regression of $y$ on $x$ is also called the coefficient of regression of $y$ on $x$. It represents the increment in the value of $y$ corresponding to a unit change in the value of $x$.

$b_{yx}$= Regression coefficient of $y$ on $x = r\frac{\sigma_x}{\sigma_y}$

## Expression for Regression Coefficient

(1) We know that

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}}$$

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$\sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

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

$$= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

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

$$= \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$$

(2) We know that

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

$$\sigma_x = \sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$\sigma_y = \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}$$

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

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

(3) We know that

$$r = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}} \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$\sigma_x = \sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}$$

$$\sigma_y = \sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_y^2 - \frac{(\sum d_y)^2}{n}}}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}}$$

## 12. Properties of Regression Coefficients

**1. The coefficient of correlation is the geometric mean of the coefficients of regression, i.e., $r = \sqrt{b_{yx} b_{xy}}$**

**Proof:** We know that

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

$$\Rightarrow b_{yx} b_{xy} = r\frac{\sigma_y}{\sigma_x}\ r\frac{\sigma_x}{\sigma_y} = r^2$$

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

**2. If one of the regression coefficients is greater than one, the other must be less than one.**

**Proof:** Let $b_{yx} > 1$

We know that

$$r^2 \leq 1 \; and \; r^2 = b_{yx} b_{xy}$$

$$b_{yx} b_{xy} \leq 1$$

$$b_{yx} \leq \frac{1}{b_{xy}}$$

Hence, if $b_{yx} < 1, b_{xy} > 1$

**3. The arithmetic mean of regression coefficients is greater than or equal to the coefficient of correlation.**

**Proof:** We have to prove that

$$\frac{1}{2}\left(b_{yx} + b_{xy}\right) \geq r$$

$$\Rightarrow \frac{1}{2}\left(r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y}\right) \geq r$$

$$\Rightarrow \frac{\sigma_y}{\sigma_x} + \frac{\sigma_x}{\sigma_y} \geq 2$$

$$\Rightarrow \sigma_y^2 + \sigma_x^2 - 2\sigma_x \sigma_y \geq 0$$

$$\Rightarrow \left(\sigma_y - \sigma_x\right)^2 \geq 0$$

Which is always true, since the square of a real quantity is $1 \geq 0$

**4.Regression coefficient are independent of the change of origin but not of scale.**

**Proof:** Let $d_x = \frac{x-a}{h}, d_y = \frac{y-b}{k}$

$$\Rightarrow x = a + hd_x, y = b + kd_y$$

Where $a, b, h(> 0)$ and $k(> 0)$ are constants.

$$r_{d_x d_y} = r_{xy}, \sigma_{d_x}^2 = \frac{1}{h^2}\sigma_x^2, \sigma_{d_y}^2 = \frac{1}{k^2}\sigma_y^2$$

$$b_{d_x d_y} = r_{d_x d_y}\frac{\sigma_{d_x}}{\sigma_{d_y}}$$

$$= r_{xy}\frac{\sigma_x}{h}\frac{k}{\sigma_y}$$

$$= \frac{k}{h}\, r_{xy}\frac{\sigma_x}{\sigma_y}$$

$$= \frac{k}{h}\, b_{xy}$$

Similarly, $b_{d_y d_x} = \frac{h}{k} b_{yx}$

**5. Both regression coefficients will have the same sign i.e., either both are positive or both are negative.**

**6. The sign of correlation is same as that of the regression coefficients, i.e., $r > 0$ if $b_{xy} > 0$ and $b_{yx} > 0$; and $r < 0$ if $b_{xy} < 0$ and $b_{yx} < 0$.**

## 13. Properties of Linear Regression

1. The two regression lines $x$ on $y$ and $y$ on $x$ always intersect at their means $(\bar{x}, \bar{y})$.

2.Since $r^2 = b_{yx} b_{xy}$, i.e., $r = \sqrt{b_{yx} b_{xy}}$, therefore $r, b_{yx}, b_{xy}$ all have the same sign.

3. If $r = 0$, the regression coefficients are zero.

4. The regression lines become identical if $r = \pm 1$. It follows from the regression equation that $x = \bar{x}$ and $y = \bar{y}$. If $r = 0$, these lines are perpendicular to each other.

**Example 11: The regression lines of a sample are $x + 6y = 6$ and $3x + 2y = 10$. Find**

**(1) sample means $\bar{x}$ and $\bar{y}$, and**

**(2) the coefficient of correlation between $x$ and $y$.**

**(3) Also estimate $y$ when $x = 12$.**

**Solution:**

(1) The regression lines pass through the point $(\bar{x}, \bar{y})$

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

$$3\bar{x} + 2\bar{y} = 10$$

Solving above equations we get

$$\bar{x} = 3, \bar{y} = \frac{1}{2}$$

(2) Let the line $x + 6y = 6$ be the line of regression of $y$ on $x$.

$$6y = -x + 6$$

$$y = -\frac{1}{6}x + 1$$

$$\therefore b_{yx} = -\frac{1}{6}$$

Let the line $3x + 2y = 10$ be the line of regression of $x$ on $y$.

$$3x = -2y + 10$$

$$x = -\frac{2}{3}y + \frac{10}{3}$$

$$b_{xy} = -\frac{2}{3}$$

Now, $r = \sqrt{b_{yx}b_{xy}} = \sqrt{\left(-\frac{1}{6}\right)\left(-\frac{2}{3}\right)} = \frac{1}{3}$

Since $b_{yx}$ and $b_{xy}$ are negative, $r$ is negative.

$$r = -\frac{1}{3}$$

Estimated value of $y$ when $x = 12$ is

$$y = -\frac{1}{6}(12) + 1 = -1$$

**<u>Example 12</u>: The following data regarding the height (y) and weight (x) of 100 college students are given:**

$$\boldsymbol{\sum x = 15000, \sum x^2 = 2272500, \sum y = 6800, \sum y^2 = 463025, \sum xy = 1022250}$$

**Find the coefficient of correlation between height and weight and also the equation of regression of height and weight.**

**Solution**

$n = 100$

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

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{2272500 - \frac{(15000)^2}{100}}$$

$= 0.1$

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

$$= \frac{1022250 - \frac{(15000)(6800)}{100}}{463025 - \frac{(6800)^2}{100}}$$

$= 3.6$

$r = \sqrt{b_{yx} b_{xy}} = \sqrt{(0.1)(3.6)} = 0.6$

$\bar{x} = \frac{\sum x}{n} = \frac{15000}{100} = 150$

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

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

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

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

$y = 0.1x + 53$

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

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

$x - 150 = 3.6(y - 68)$

$x = 3.6y - 94.8$

**<u>Example 13</u>: Find the regression coefficients $b_{yx}$ and $b_{xy}$ and hence, find the correlation coefficient between $x$ and $y$ for the following data:**

| $x$ | 4 | 2 | 3 | 4 | 2 |
|---|---|---|---|---|---|
| $y$ | 2 | 3 | 2 | 4 | 4 |

**Solution:**

$n = 5$

| $x$ | $y$ | $x^2$ | $y^2$ | $xy$ |
|---|---|---|---|---|
| 4 | 2 | 16 | 4 | 8 |
| 2 | 3 | 4 | 9 | 6 |
| 3 | 2 | 9 | 4 | 6 |
| 4 | 4 | 16 | 16 | 16 |
| 2 | 4 | 4 | 16 | 8 |
| $\sum x = 15$ | $\sum y = 15$ | $\sum x^2 = 49$ | $\sum y^2 = 49$ | $\sum xy = 44$ |

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

$$= \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}}$$

$$= -0.25$$

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

$$= \frac{44 - \frac{(15)(15)}{5}}{49 - \frac{(15)^2}{5}}$$

$$= -0.25$$

$$r = \sqrt{b_{yx} b_{xy}} = \sqrt{(-0.25)(-0.25)} = 0.25$$

Since $b_{yx}$ and $b_{xy}$ are negative, $r$ is negative.

$$r = -0.25$$

**Example-14: The number of bacterial cells $(y)$ per unit volume in a culture at different hours $(x)$ is given below:**

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 43 | 46 | 82 | 98 | 123 | 167 | 199 | 213 | 245 | 272 |

**Fit lines of regression of $y$ on $x$ and $x$ on $y$. Also, estimate the number of bacterial cells after 15 hours.**

**Solution:**

| $x$ | $y$ | $x^2$ | $xy$ | $y^2$ |
|---|---|---|---|---|
| 0 | 43 | 0 | 0 | 1849 |
| 1 | 46 | 1 | 46 | 2116 |
| 2 | 82 | 4 | 164 | 6742 |
| 3 | 98 | 9 | 294 | 9604 |
| 4 | 123 | 16 | 492 | 15129 |
| 5 | 167 | 25 | 835 | 27889 |
| 6 | 199 | 36 | 1194 | 36801 |
| 7 | 213 | 49 | 1491 | 45369 |
| 8 | 245 | 68 | 1960 | 60025 |
| 9 | 272 | 81 | 2448 | 73984 |
| $\sum x = 45$ | $\sum y = 1488$ | $\sum x^2 = 285$ | $\sum xy = 8924$ | $\sum y^2 = 282290$ |

$n = 10$

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

$$= \frac{8924 - \frac{(48)(1488)}{10}}{285 - \frac{(45)^2}{10}}$$

$$= 27.0061$$

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

$$= \frac{8924 - \frac{(45)(1488)}{10}}{282290 - \frac{(1488)^2}{10}}$$

$$= 0.0366$$

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{10} = 4.5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{1488}{10} = 148.8$$

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

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

$$y - 148.8 = 27.0061(x - 4.5)$$

$$y = 27.0061x + 27.2726$$

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

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

$$x - 4.5 = 0.0366(y - 148.8)$$

$$x = 0.0366y - 0.9461$$

At $x = 15$ hours,

$$y = 27.0061(15) + 27.2726 = 432.3641$$

**<u>Example-15</u>: Find the two lines of regression from the following data:**

| **Age of husband(x)** | **25** | **22** | **28** | **26** | **35** | **20** | **22** | **0** | **20** | **18** |
|---|---|---|---|---|---|---|---|---|---|---|
| **Age of wife (y)** | **18** | **15** | **20** | **17** | **22** | **14** | **16** | **21** | **15** | **14** |

**Hence, estimate $(i)$ the age of the husband when the age of the wife is 19, and $(ii)$ the age of the wife when the age of the husband is 30.**

**Solution:**

Let $a = 26$ and $b = 17$ be the assumed means of $x$ and $y$ series respectively.

$d_x = x - a = x - 26$

$d_y = y - b = y - 17$

$n = 10$

| $x$ | $y$ | $d_x$ | $d_y$ | $d_x^2$ | $d_y^2$ | $d_x d_y$ |
|---|---|---|---|---|---|---|
| 25 | 18 | -1 | 1 | 1 | 1 | -1 |
| 22 | 15 | -4 | -2 | 16 | 4 | 8 |
| 28 | 20 | 2 | 3 | 4 | 9 | 6 |
| 26 | 17 | 0 | 0 | 0 | 0 | 0 |
| 35 | 22 | 9 | 5 | 81 | 25 | 45 |
| 20 | 14 | -6 | -3 | 36 | 9 | 18 |
| 22 | 16 | -4 | -1 | 16 | 1 | 4 |
| 40 | 21 | 14 | 4 | 196 | 16 | 56 |
| 20 | 15 | -6 | -2 | 36 | 4 | 12 |
| 18 | 14 | -8 | -3 | 64 | 9 | 24 |
| $\sum x = 256$ | $\sum y = 172$ | $\sum d_x = -4$ | $\sum d_y = 2$ | $\sum d_x^2 = 450$ | $\sum d_y^2 = 78$ | $\sum d_x d_y = 172$ |

$$b_{xy} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_y^2 - \frac{\left(\sum d_y\right)^2}{n}}}$$

$$= \frac{172 - \frac{(-4)(2)}{10}}{450 - \frac{(-4)^2}{10}}$$

$$= 0.385$$

$$b_{yx} = \frac{\sum d_x d_y - \frac{\sum d_x \sum d_y}{n}}{\sqrt{\sum d_x^2 - \frac{(\sum d_x)^2}{n}}}$$

$$= \frac{172 - \frac{(-4)(2)}{10}}{78 - \frac{(2)^3}{10}}$$

$$= 2.227$$

$\bar{x} = \frac{\sum x}{n} = \frac{256}{10} = 25.6$

$\bar{y} = \frac{\sum y}{n} = \frac{172}{10} = 17.2$

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

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

$y - 17.2 = 0.385(x - 25.6)$

$y = 0.385x + 7.344$

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

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

$x - 25.6 = 2.227(y - 17.2)$

$x = 2.227y - 12.704$

Estimated age of the husband when the age of the wife is 19 is

$x = 2.227(19) - 12.704 = 29.601\ or\ 30$ nearly

Age of the husband=30 years

Estimated age of the wife when the age of the husband is 30 is

$y = 0.385(30) + 7.344 = 18.894\ or\ 19$ nearly

Age of the wife = 19 years