Chapter 5

Correlation, Regression, Time Series Analysis and Index Numbers

5.1 Correlation and Regression

Correlation analysis is the statistical tool used to measure the degree to which two variables are linearly related to each other. Correlation measures the degree of association between two variables.

If the quantities (X, Y) vary in such a way that change in one variable corresponds to change in the other variable, then the variables X and Y are correlated.

Example: Price of commodity and amount of demand.

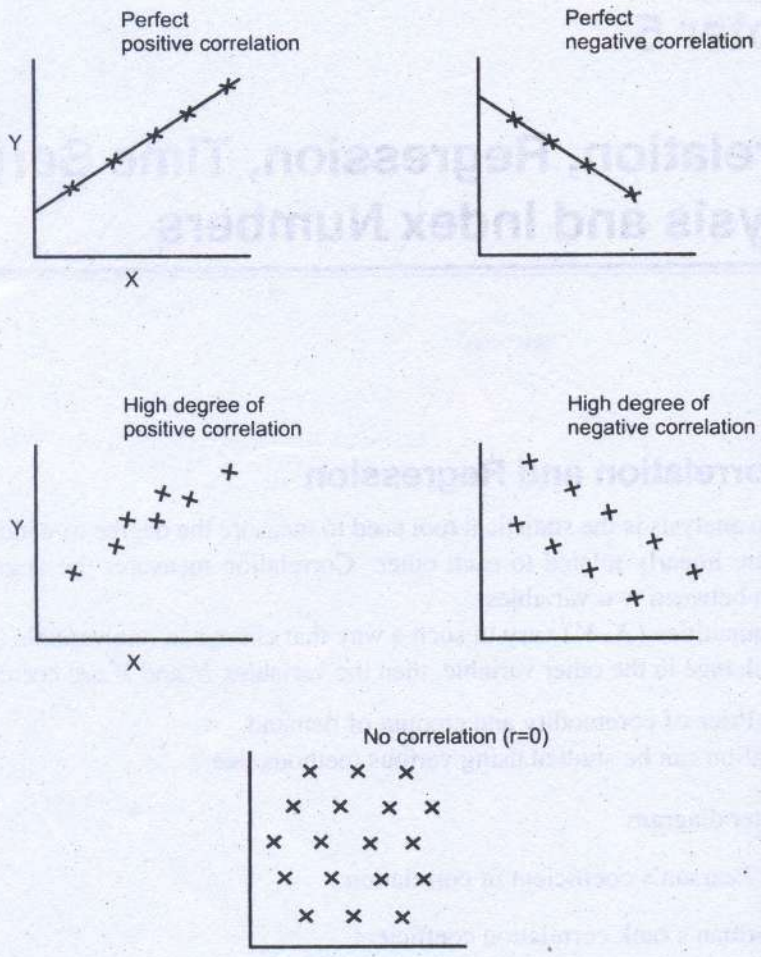
Correlation can be studied using various methods like

1. Scatter diagram
2. Karl Pearson's coefficient of correlation
3. Spearman's rank correlation coefficient.

5.1.1 Scatter Diagram

The simplest device for studying correlation between two variables is a special type of dot chart called scatter diagram. When this method is used, the given data are plotted on a graph in the form of dots.

i.e, for each pair of X and Y values we put dots and thus we obtain as many points as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the 2 variables are related or not. The more the plotted points scatter over a chart, the lesser is the degree of relationship between the two variables. The nearer the points come to the line, the higher the degree of relationship. If the plotted points lie in a haphazard manner it shows the absence of any relationship between the variables. Consider the following diagrams.



5.1.2 Karl Pearson's co-efficient of correlation (Product moment correlation co-efficient)

The co-efficient of correlation between X and Y is defined as

$$r(X,Y) = \frac{\text{Cov}\,(X,Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$$

$$= \frac{\text{Cov}\,(X,Y)}{\sigma_x\,\sigma_y}$$

$$= \frac{E\left[(X-\overline{X})(Y-\overline{Y})\right]}{E[X-\overline{X}]^2\,E[Y-\overline{Y}]^2}$$

5.1.3 Properties of correlation coefficient

(i) The coefficient of correlation lies between -1 and +1 or $|r| \leq 1$.

Proof:

The correlation coefficient 'r' between X and Y is given by

$$r = \frac{\text{Cov}(x,y)}{\sigma x\, \sigma y} \quad \text{where Cov } (x,y) = \frac{1}{n}\sum(x-\overline{x})(y-\overline{y})$$

$$= \frac{K}{\sigma x\, \sigma y} \quad \text{say Cov (x,y) = K}$$

Here $K^2 = \left(\frac{1}{n}\sum(x-\overline{x})(y-\overline{y})\right)^2 = \left(\frac{\sum XY}{n}\right)^2$

$$\sigma_x^2\, \sigma_y^2 = \frac{\sum(x-\overline{x})^2}{n} \cdot \frac{\sum(y-\overline{y})^2}{n} = \frac{\sum X^2 \cdot \sum Y^2}{n^2}$$

where $X = x - \overline{x}$ and $Y = y - \overline{y}$

By Schwarz's inequality, we have

$$(\sum XY)^2 \leq (\sum X^2)(\sum Y^2)$$

Dividing both sides by n^2, we get

$$K^2 \leq \sigma_x^2 \cdot \sigma_y^2$$

$$\therefore\ r^2 \leq 1 \quad \Rightarrow \quad |r| \leq 1 \quad \text{(or)} \quad -1 \leq r \leq 1$$

Note:

1. If $r = 1$ then there is a perfect positive correlation.
2. If $r = -1$ then there is a perfect negative correlation.
3. If $r = 0$ then the variables are uncorrelated.

(ii) The co-efficient of correlation is independent of change of scale and origin of the variables X and Y.

Proof

$$\text{Let } U = \frac{X-a}{h}, \quad V = \frac{Y-b}{k}, \quad \text{so that}$$

$X = a + hU$ and $Y = b + kV$ where a, b, h, k are constants; $h > 0, k > 0$.

We shall prove that $r(X,Y) = r(U,V)$

Since $X = a + hU$ and $Y = b + kV$, on taking expectations, We have

$$E(X) = a + hE(U) \quad \text{and} \quad E(Y) = b + kE(V)$$

$$\therefore \; X - E(X) = h[U - E(U)] \quad \text{and} \quad Y - E(Y) = k[V - E(V)]$$

$$\Rightarrow \quad \text{Cov}\ (X,Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= E[h\{U - E(U)\}\ k\{V - E(V)\}]$$

$$= hkE[\{U - E(U)\}\{V - E(V)\}]$$

$$= hk\ \text{Cov}\ (U,V) \tag{1}$$

$$\sigma_X^2 = E[\{X - E(X)\}^2]$$

$$= [h^2\{U - E(U)\}^2)] = h^2\,\sigma_U^2$$

$$\Rightarrow \quad \sigma_X = h\sigma_U \quad (h > 0) \tag{2}$$

$$\text{Similarly,} \quad \sigma_Y = k\sigma_V \ (k > 0) \tag{3}$$

$$\text{We know that} \quad r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{4}$$

Substituting (1), (2) and (3) in (4), we get

$$r(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{hk\text{Cov}(U,V)}{hk\,\sigma_U\sigma_V}$$

$$= \frac{\text{Cov}(U,V)}{\sigma_U\sigma_V} = r(U,V)$$

Note: Method for finding correlation co-efficient (discrete case)

$$r = \Sigma\frac{(X - \overline{X})(Y - \overline{Y})}{n\sigma_X\sigma_Y}$$

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

5.1.4 Calculation of co-efficient of correlation for a bi-variate distribution

If the bi-variate data in x and y is given by two way table and f is the frequency, then,

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

Here also the change in origin and scale does not affect the co-efficient of correlation.

$$r_{xy} = r_{uv}$$

where u, v are new variables properly chosen.

5.1.5 Regression

Definition

Regression is the measure of the average relationship between two or more variables in terms of original units of data.

Example: If the sales and advertising are correlated we can find out the expected amount of sales for a given advertising expenditure or the amount needed for attaining the given amount of sales.

5.1.6 Lines of regression

If two variables X and Y are correlated i.e., there exists an association between them, we can see that the scatter diagram will be more or less concentrated around a curve. This curve is called ***Curve of regression***.

If the curve is a straight line, it is called the line of regression and the regression is a linear regression.

We shall have two regression lines as the regression line of X and Y and the regression line of Y and X. The regression line of Y and X gives the most probable value of Y for given values of X and the regression line of X and Y gives the most probable values of X for given values of Y.

Table 5.1 Relation between Correlation Analysis and Regression Analysis

S.No	Correlation Analysis	Regression Analysis
1.	Correlation coefficient r between X and Y is a measure of linear relationship between X and Y	The regression coefficients are mathematical measures expressing the average relationship between the two variables.
2.	The correlation coefficient does not reflect upon the nature of variable (independent or dependent variable)	Regression coefficients reflect on the nature of variable i.e, which is dependent variable. In other words, it estimates the value of dependent variable for any given value of independent variable.

3.	It does not imply cause and effect relationship between the variables under study	It indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable, whereas corresponding to effect is taken as dependent variable.
4.	It is a relative measure and is independent of the units of measurement	Regression coefficients are absolute measures of finding out the relationship between two or more variables
5.	It indicates the degree of association.	It is used to forecast the nature of dependent variable when the value of independent variable is known.

Uses of Regression Analysis

1. The cause and effect relations are indicated from the study of regression analysis.
2. It establishes the rate of change in one variable in terms of the changes in another variable.
3. It is useful in economic analysis as regression equation can determine an increase in the cost of living index for a particular increase in general price level.
4. It helps in prediction and thus it can estimate the value of unknown quantities.
5. It enables us to study the nature of relationship between the variables.
6. It can be useful to all natural, social and physical sciences, where the data are in functional relationship.

5.1.7 Regression Equations

(i) Equation of line of regression of Y on X is

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

where $r \frac{\sigma_y}{\sigma_x}$ is the regression coefficient of Y on X.

(ii) Equation of line of regression of X on Y is

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

where $r \frac{\sigma_x}{\sigma_y}$ is the regression co-efficient of X on Y.

Note:

1. The regression coefficients can be denoted by

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

2. The regression co-efficients are obtained by the following expressions for discrete values of X and Y

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

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

3. Both the regression lines pass through the point $(\overline{x}, \overline{y})$ where $\overline{x}$ and $\overline{y}$ are means of X and Y respectively.
4. Correlation coefficient is the Geometric mean between the regression coefficients.

$$\text{i.e,}\quad b_{xy} \cdot b_{yx} = r^2 \;\Rightarrow\; r = \pm\sqrt{b_{xy} \cdot b_{yx}}$$

5. If one of the regression coefficients is greater than unity the other must be less than unity.
6. Regression coefficients are independent of the change of origin but not of scale.
7. Both the regression coefficients will have the same sign, i.e, they will be either both positive or both negative. The coefficient correlation will have the same sign as that of regression coefficients, i.e, if regression coefficients have a negative sign, r will also have negative sign and if the regression coefficients have a positive sign, r will also be positive.

5.1.8 Angle between regression lines

If θ is the angle between the two regression lines, then

$$tan\,\theta = \left(\frac{1 - r^2}{r}\right)\frac{\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2}$$

where $r, \sigma x, \sigma y$ have the usual meaning.

Proof :

Equation of the regression lines of Y on X and X on Y are

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

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