Chapter 6

Correlation Analysis

Prepared by ***Jannatul Ferdous***, Assistant Professor, Department of Computer Science and Engineering

Correlation Analysis

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5 ft. 5 inch. is less than the average height of people 5 ft. 6 inch., and their average weight is less than that of people 5 ft. 7 inch., etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.

Definition: The **Correlation Analysis** is the statistical tool used to study the closeness of the relationship between two or more variables. The variables are said to be correlated when the movement of one variable is accompanied/changed by the movement of another variable.

The correlation analysis is used when the researcher wants to determine the possible association between the variables and to begin with; the following steps are to be followed:

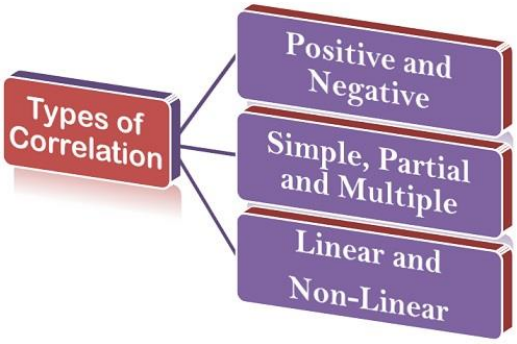
- Determining whether the relation exists and then measuring it (The measure of correlation is called as the **Coefficient of Correlation**).
- Testing its significance
- Establishing the cause-and-effect relation, if any.

In the correlation analysis, there are two types of variables- **Dependent and Independent**. The purpose of such analysis is to find out if any change in the independent variable results in the change in the dependent variable or not. Now the question arises that what is the need to study the correlation? The study of correlation is very useful in the practical life due to the following reasons:

1. Several variables show some kind of relationship, such as income and expenditure, demand and sales, etc. and hence, with the help of correlation analysis the degree of relationship between these variables can be measured in one figure.
2. Once the closeness of variables is determined, we can estimate the value of unknown variable provided the value of another variable is given. This can be done using the regression analysis.
3. The correlation analysis helps the manufacturing firm in estimating the price, cost, sales of its product on the basis of the other variables that are functionally related to it.
4. It contributes towards the economic behavior as it helps an economist in identifying the critically important variables on which several other economic variables depend on.

The correlation analysis is the most widely used method and is often the most abused statistical measures. This is because the researcher may overlook the fact that the correlation only measures the **strength of linear relationships** and does not necessarily imply a relationship between the variables.

Types of Correlation



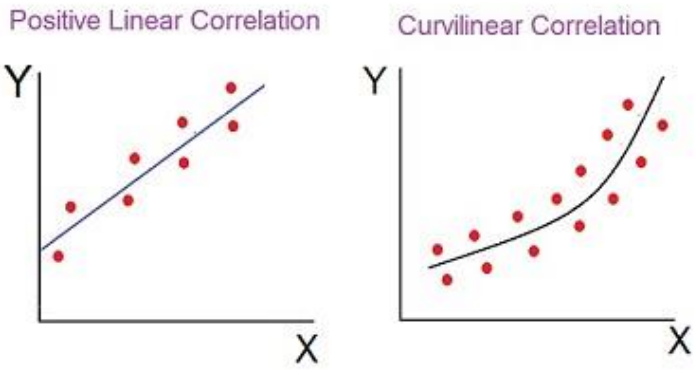
1. **Positive and Negative Correlation:** Whether the correlation between the variables is positive or negative depends on its **direction of change**. The correlation is positive when both the variables **move in the same direction**, i.e. when one variable increases the other on an average also increases and if one variable decreases the other also decreases. The correlation is said to be negative when both the variables **move in the opposite direction**, i.e. when one variable increases the other decreases and vice versa.
2. **Simple, Partial and Multiple Correlation:** Whether the correlation is simple, partial or multiple depends on the **number of variables studied**. The correlation is said to be simple when **only two variables** are studied. The correlation is either multiple or partial when three or more variables are studied. The correlation is said to be Multiple when **three variables are studied simultaneously**. Such as, if we want to study the relationship between the yield of wheat per acre and the amount of fertilizers and rainfall used, then it is a problem of multiple correlations.

   Whereas, in the case of a partial correlation we study more than two variables, **but consider only two among them that would be influencing each** other such that the effect of the other influencing variable is kept constant. Such as, in the above example, if we study the relationship between the yield and fertilizers used during the periods when certain average temperature existed, then it is a problem of partial correlation.
3. **Linear and Non-Linear (Curvilinear) Correlation:** Whether the correlation between the variables is linear or non-linear depends on the **constancy of ratio of change between the variables**. The correlation is said to be linear when the amount of change in one variable to the amount of change in another variable tends to **bear a constant ratio**. For example, from the values of two variables given below, it is clear that the ratio of change between the variables is the same:

X:	**10**	**20**	**30**	**40**	**50**
Y:	**20**	**40**	**60**	**80**	**100**

   The correlation is called as non-linear or curvilinear when the amount of change in one variable **does not bear a constant ratio** to the amount of change in the other variable. For example, if the amount of fertilizers is doubled the yield of wheat would not be necessarily be doubled.

Thus, these are three most important types of correlation classified on the basis of movement, number and the ratio of change between the variables. The researcher must study these carefully to determine the correlation methods to be used to identify the extent to which the variables are correlated.



Methods of Determining Correlation

These figures clearly show the difference between the linear and non-linear correlation. To determine the linearity and non-linearity among the variables and the extent to which these are correlated, following are the important methods used to ascertain these:

1. Scatter Diagram Method
2. Karl Pearson's Coefficient of correlation
3. Spearman's Rank Correlation Coefficient
4. Methods of Least Squares

Among these, the first method, i.e. scatter diagram method is based on the study of graphs while the rest is mathematical methods that use formulae to calculate the degree of correlation between the variables. The researchers may apply either of these methods on the basis of the nature of variables being considered in ascertaining the association between them.

Karl Pearson's Coefficient of Correlation

Definition: Karl Pearson's Coefficient of Correlation is widely used mathematical method where in the numerical expression is used to calculate the degree and direction of the relationship between linear related variables.

Pearson's method, popularly known as a **Pearsonian Coefficient of Correlation,** is the most extensively used quantitative methods in practice. The coefficient of correlation is denoted by **"r".**

If the relationship between two variables X and Y is to be ascertained, then the following formula is used:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson's correlation coefficient between x and y
n = number of observations
x_i = value of x (for i^{th} observation)
y_i = value of y (for i^{th} observation)

The correlation coefficient, denoted by r, is a measure of the strength of the straight-line or linear relationship between two variables. The correlation coefficient takes on values ranging between +1 and -1. The following points are the accepted guidelines for interpreting the correlation coefficient:

1. 0 indicates no linear relationship.
2. +1 indicates a perfect positive linear relationship: as one variable increases in its values, the other variable also increases in its values via an exact linear rule.
3. -1 indicates a perfect negative linear relationship: as one variable increases in its values, the other variable decreases in its values via an exact linear rule.
4. Values between 0 and 0.3 (0 and -0.3) indicate a weak positive (negative) linear relationship.
5. Values between 0.3 and 0.7 (-0.3 and -0.7) indicate a moderate positive (negative) linear relationship.
6. Values between 0.7 and 1.0 (-0.7 and -1.0) indicate a strong positive (negative) linear relationship.

Properties of Coefficient of Correlation

- The value of the coefficient of correlation (r) always **lies between ±1**. Such as:
  r = +1,perfectpositive correlation
  r = -1, perfect negative correlation
  r = 0, no correlation
- The coefficient of correlation is independent of the **origin and scale.** By origin, it means subtracting any non-zero constant from the given value of X and Y the value of "r" remains unchanged. By scale it means, there is no effect on the value of "r" if the value of X and Y is divided or multiplied by any constant.

The coefficient of correlation is a **geometric mean of two regression coefficient.** Symbolically it is represented as: $r = \sqrt{b_{xy} \times b_{yx}}$

- The coefficient of correlation is **"zero"** when the variables X and Y are independent. But, however, the converse is not true.

Assumptions of Karl Pearson's Coefficient of Correlation r

1. The relationship between the variables is **"Linear",** which means when the two variables are plotted, a straight line is formed by the points plotted.
2. There are a large number of independent causes that affect the variables under study so as to form a **Normal Distribution**. Such as, variables like price, demand, supply, etc. are affected by such factors that the normal distribution is formed.
3. The variables are independent of each other.

Note: The coefficient of correlation measures not only the magnitude of correlation but also tells the direction. Such as, r = -0.67, which shows correlation is negative because the sign is **' — '** and the magnitude is **0.67**.

Example 1: Find the value of the correlation coefficient from the following table:

Subject	1	2	3	4	5	6
Age (x)	43	21	25	42	57	59
Glucose Level (y)	99	65	79	75	87	81

Solution 1:

Subject	**Age x**	**Glucose Level y**	**xy**	$\mathbf{x^2}$	$\mathbf{y^2}$
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
n = 6	Σ x =247	Σy = 486	Σ xy = 20485	Σ $\mathbf{x^2}$ = 11409	Σ $\mathbf{y^2}$ = 40022

Use the following correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2]\,[\, n\Sigma y^2 - (\Sigma y)^2]}}$$

From the table:

$\Sigma x = 247$

$\Sigma y = 486$

$\Sigma xy = 20485$

$\Sigma x^2 = 11409$

$\Sigma y^2 = 40022$

$n = 6$

The correlation coefficient,

$r = 6(20,485) - (247 \times 486) / [\sqrt{[[6(11,409) - (247^2)]} \times [6(40,022) - 486^2]]]$

$= 2868 / 5413.27 = 0.529809$

Comment/ Interpretation: The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.