SNS COLLEGE OF ENGINEERING

Kurumbapalayam (po), Coimbatore – 641 107

Accredited by NAAC-UGC with 'A' Grade

Approved by AICTE & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

19AD504 – DATA VISUALIZATION

UNIT - 4

Correlation and Regression

- Correlation and regression are the two most commonly used techniques for investigating the relationship between quantitative variables.
- Here regression refers to linear regression. Correlation is used to give the relationship between the variables whereas linear regression uses an equation to express this relationship.
- Correlation and regression are used to define some form of association between quantitative variables that are assumed to have a linear relationship.

What are Correlation and Regression?

- Correlation and regression are statistical measurements that are used to give a relationship between two variables.
- For example, suppose a person is driving an expensive car then it is assumed that she must be financially well. To numerically quantify this relationship, correlation and regression are used.

Correlation Definition

- Correlation can be defined as a measurement that is used to quantify the relationship between variables.
- If an increase (or decrease) in one variable causes a corresponding increase (or decrease) in another then the two variables are said to be directly correlated.
- Similarly, if an increase in one causes a decrease in another or vice versa, then the variables are said to be indirectly correlated.

- If a change in an independent variable does not cause a change in the dependent variable then they are uncorrelated.
- Thus, correlation can be positive (direct correlation), negative (indirect correlation), or zero. This relationship is given by the correlation coefficient.

Regression Definition

- Regression can be defined as a measurement that is used to quantify how the change in one variable will affect another variable.
- Regression is used to find the cause and effect between two variables.
- Linear regression is the most commonly used type of regression because it is easier to analyze as compared to the rest.
- Linear regression is used to find the line that is the best fit to establish a relationship between variables.

Correlation and Regression Analysis

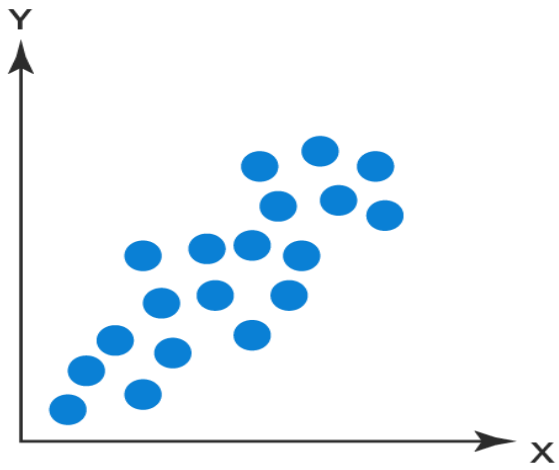
- Both correlation and regression analysis are done to quantify the strength of the relationship between two variables by using numbers.
- Graphically, correlation and regression analysis can be visualized using scatter plots.

Correlation analysis is done so as to determine whether there is a relationship between the variables that are being tested.

Furthermore, a correlation coefficient such as Pearson's correlation coefficient is used to give a signed numeric value that depicts the strength as well as the direction of the correlation.

The scatter plot gives the correlation between two variables x and y for individual data points as shown below.

Correlation Analysis Graph

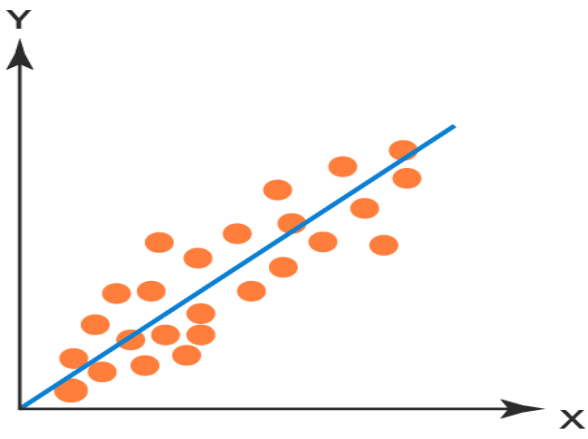


Regression analysis is used to determine the relationship between two variables such that the value of the unknown variable can be estimated using the knowledge of the known variables.

The goal of linear regression is to find the best-fitted line through the data points.

For two variables, x, and y, the regression analysis can be visualized as follows:

Regression Analysis Graph



Difference between Correlation and Regression

Correlation and regression are both used as statistical measurements to get a good understanding of the relationship between variables.

If the correlation coefficient is negative (or positive) then the slope of the regression line will also be negative (or positive).

The table given below highlights the key difference between correlation and regression.

Correlation	**Regression**
Correlation is used to determine whether variables are related or not.	Regression is used to numerically describe how a dependent variable changes with a change in an independent variable
Correlation tries to establish a linear relationship between variables.	It finds the best-fitted regression line to estimate an unknown variable on the basis of the known variable.
The variables can be used interchangeably	The variables cannot be interchanged.
Correlation uses a signed numerical value to estimate the strength of the relationship between the variables.	Regression is used to show the impact of a unit change in the independent variable on the dependent variable.
The Pearson's coefficient is the best measure of correlation.	The least-squares method is the best technique to determine the regression line.