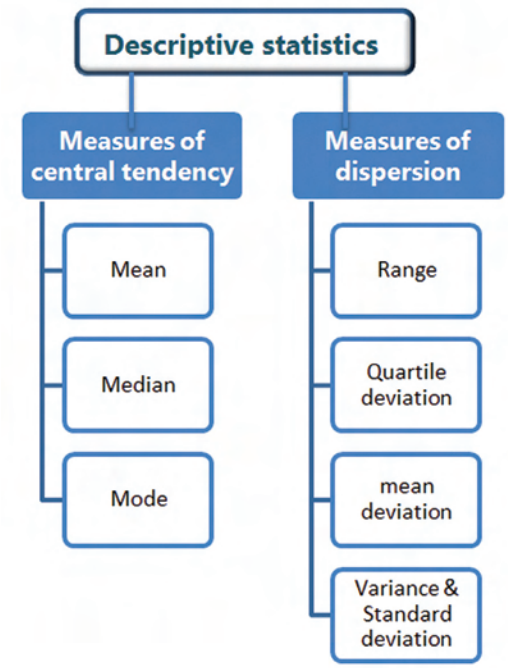
CHAPTER
10

STATISTICAL METHODS

Somy Kuriakose

Fishery Resources Assessment Division
ICAR-Central Marine Fisheries Research Institute

Statistics plays a central role in research, planning and decision-making in almost all natural and social sciences. It is the Science of collecting, organizing, analyzing, interpreting and presenting data. It deals with all aspects of this, including not only the collection, analysis and interpretation of such data, but also the planning the collection of data, in terms of the design of surveys and experiments. Two types of statistical methods are used in analysing data: descriptive statistics and inferential statistics. **Inferential statistics** makes inferences and predictions about a population based on a sample of data taken from the population in question. **Descriptive statistics** uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables. Descriptive statistics therefore enables us to present the data in a more meaningful way, which allows simpler interpretation of the data.



Measures of central tendency

Description of a variable usually begins with the specification of its single most representative value, often called the measure of central tendency. The best way to reduce a set of data and still retain part of the information is to summarize the set with a

single value. A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. Measures of central tendency are sometimes called measures of central location or summary statistics. Measures of central tendency are measures of the location of the middle or the center of a distribution. There are several measures for this statistic.

Measures of central tendency

Arithmetic mean

The arithmetic mean of a set of values is the quantity commonly called the mean or the average. For a data set, the mean is the sum of the values divided by the number of values. The mean of a set of numbers $x_1, x_2, ...,x_n$ is typically denoted by $\bar{x}$ pronounced "x bar".

Arithmetic Mean = $\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$ Or $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Arithmetic Mean from a grouped data

i) Discrete frequency distribution

Data arising from organising 'n' observed values into a smaller number of disjoint groups of values, and counting the frequency of each group; often presented as a frequency table. In this case the values of the variable are multiplied by their respective frequencies and this total is then divided by the total number of frequencies.

Arithmetic mean, $\bar{x} = \frac{f_1 x_1 + f_2 x_2 + ... + f_n x_n}{f_1 + f_2 + ... + f_n}$ $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$

where $x_1, x_2, ... x_n$ are values of the variable x and $f_1, f_2, ...f_n$ are their corresponding frequencies.

ii) Continuous frequency distribution

We take mid values of each class as representative of that class, multiply this mid values by their corresponding frequencies, total these products and divide by the total number of items. If $x_1, x_2, ...x_n$ represent the mid values of classes and $f_1, f_2, ...f_n$ the frequencies, then

$$\text{Arithmetic Mean} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots\ldots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

$$\text{Where } N = \sum_{i=1}^{i=n} f_i$$

The mean is valid only for interval data or ratio data. Since it uses the values of all of the data points in the population or sample, the mean is influenced by outliers that may be at the extremes of the data set. The mean uses all the observations and each observation affects the mean. Even though the mean is sensitive to extreme values (*i.e.*, extremely large or small data can cause the mean to be pulled toward the extreme data) it is still the most widely used measure of location. This is due to the fact that the mean has valuable mathematical properties that make it convenient for use with inferential statistics analysis. For example, the sum of the deviations of the numbers in a set of data from the mean is zero, and the sum of the squared deviations of the numbers in a set of data from the mean is minimum value. The following are the merits and demerits of arithmetic mean.

Merits and Demerits of Arithmetic Mean

Merits	Demerits
•It is rigidly defined. •It is easy to calculate and simple to follow. •It is based on all the observations. •It is determined for almost every kind of data. •It is finite and not indefinite. •It is readily put to algebraic treatment. •It is least affected by fluctuations of sampling. •It is easy to calculate	•The arithmetic mean is highly affected by extreme values. •It is not an approprite average for highly skewed distributions. •It cannot be computed accurately if any item is missing.

Median

Median is the value in the middle of the data set, when the data points are arranged from smallest to largest. If there are an odd number of data points, then just arrange them in ascending or descending order and take the middle value. If there is an even number of data points, you will need to take the average of the two middle values. Hence median is determined by sorting the data set from lowest to highest values and taking the data point in the middle of the sequence. There is an equal number of points above and below the median.

Calculation of median in a grouped data

i) Discrete series

In this case also, data should be arranged in ascending or descending order of magnitude and find out the cumulative frequencies. Now find out the value of $(n+1/2)^{th}$ item. It can be found by first locating the cumulative frequency which is equal to (n+1/2) and then determine the value corresponding to it. This will be the value of median.

ii) Continuous series

For computing the value of the median in a continuous series, first determine the particular class in which the value of the median lies. Use N/2 as the rank of Median where N= total frequency. Hence it is N/2 which will divide the area of the curve into two parts. The following formula is used for determining the exact value of the median.

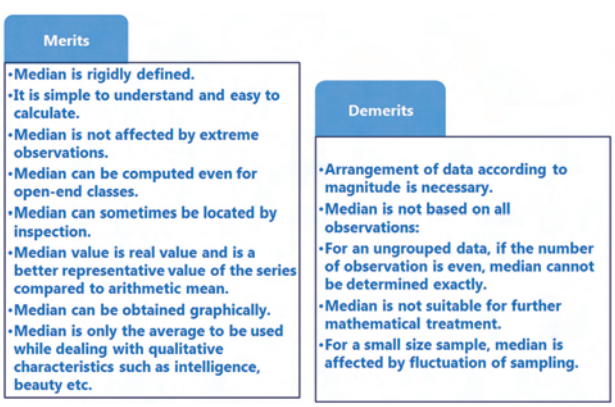
$$\text{Median} = l + \frac{(\frac{N}{2} - m) * c}{f}$$

where N= $\sum fi$ = Total frequency, l- the lower limit of the median class, m -cumulative frequency up to the median class, f- frequency of the median class and c- class width.

Median

The median can be determined for ordinal data as well as interval and ratio data. Unlike the mean, the median is not influenced by outliers at the extremes of the data set. Generally, the median provides a better measure of location than the mean when there are some extremely large or small observations (*i.e.,* when the data are skewed to the right or to the left). For this reason, the median is often used when there are a few extreme values that could greatly influence the mean and distort what might be considered typical. Note that if the median is less than the mean, the data set is skewed to the right. If the median is greater than the mean, the data set is skewed to the left. Median does not have important mathematical properties for use in future calculations.

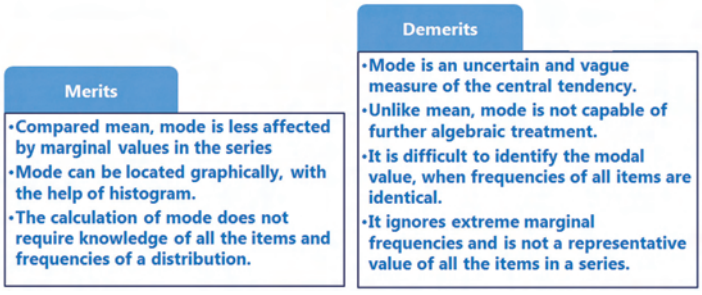
Merits and Demerits of Median



Mode

Mode is the most common value or most frequently occurring value in the data set. For finding the mode, just look at the data, count how many of each value you have, and select the data point that shows up the most frequently. If no value occurs more than once, then there is no mode. If two values occur as frequently as each other and more frequently than any other, then there are two modes. In the same way, there could also be more than two modes.

Merits and Demerits of Mode



Mode is very simple measure of central tendency. Because of its simplicity, it is a very popular measure of the central tendency. The mode can be very useful for dealing with categorical data. The mode also can be used with ordinal, interval, and ratio data. However, in interval and ratio scales, the data may be spread thinly with no data points having the same value. In such cases, the mode may not exist or may not be very meaningful. Following are the various merits and demerits of mode:

Weighted Mean

When two or more means are combined to develop an aggregate mean, the influence of each mean must be weighted by the number of cases in its subgroup.

$$\overline{X}_w = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3}{n_1 + n_2 + n_3}$$

Geometric mean (GM)

The geometric mean is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean) e.g. rates of growth.

$$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$$

Harmonic mean (HM)

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

$$\bar{x} = n \cdot \left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$$

Relationship between AM, GM, and HM

AM, GM, and HM satisfy these inequalities:

$$AM > GM > HM$$

Equality holds only when all the elements of the given sample are equal.

The mean (often called the average) is most common measure of central tendency, but there are others, such as, the median and the mode. The mean, median and mode are all valid measures of central tendency but, under different conditions, some measures of central tendency become more appropriate to use than others.

Measures of Dispersion

Measure of variation describes how spread out or scattered a set of data. It is also known as measures of dispersion or measures of spread. Measures of variation determine the range of the distribution, relative to the measures of central tendency. Measures of average such as the mean and median represent the typical value for a dataset. Within the dataset the actual values usually differ from one another and from the average value itself. The extent to which the mean and median are good representatives of the values in the original dataset depends upon the variability or dispersion in the original data. The measures of central tendency are specific data points, measures of variation are lengths between various points within the distribution. It provide us with a summary of how much the points in our data set vary, e.g. how spread out they are or how volatile they are. Measures of variation together with measures of central tendency are important for identifying key features of a sample to better understand the population from which the sample comes from. Datasets are said to have high dispersion when they contain values considerably higher and lower than the mean value. The most common measures of variation are Range, Quartile déviation or semi Interquartile Range, Mean deviation, Variance, Standard deviation and Coefficient of Variation.

Range

The range is the distance between the lowest data point and the highest data point. In other words, it is difference between the highest value and the lowest value.

Range = Highest value–lowest value

The range is the simplest measure of variation to find. Since the range only uses the largest and smallest values, it is greatly affected by extreme values, that is - it is not resistant to change.

The range is simple to compute and is useful when you wish to evaluate the whole of a dataset. It is useful for showing the spread within a dataset and for comparing the spread between similar datasets.

Since the range is based solely on the two most extreme values within the dataset, if one of these is either exceptionally high or low (sometimes referred to as outlier) it will result in a range that is not typical of the variability within the dataset. The range does not really indicate how the scores are concentrated along the distribution. The range only involves the smallest and largest numbers, and is affected by extreme data values or outliers. In order to reduce the problems caused by outliers in a dataset, the inter-quartile range is often calculated instead of the range.

Quartile Deviation or The semi inter-quartile Range

The inter-quartile range is a measure that indicates the extent to which the central 50% of values within the dataset are dispersed. If the sample is ranked in ascending order of magnitude two values of *x* may be found, the first of which is exceeded by 75% of the sample, the second by 25%; their difference is the interquartile range. It is based upon and related to the median. In the same way that the median divides a dataset into two halves, it can be further divided into quarters by identifying the upper and lower quartiles. The lower quartile, Q1 is found one quarter of the way along a dataset when the values have been arranged in order of magnitude; the upper quartile Q3 is found three quarters along the dataset. Therefore, the upper quartile lies half way between the median and the highest value in the dataset whilst the lower quartile lies halfway between the median and the lowest value in the dataset. Between Q1 and Q3 there is half the total number of items. Q3-Q1 affords a convenient and often a good indicator of the absolute variability. Usually one half of the Q3-Q1 is used and given the name semi-interquartile range or quartile deviation.

$$Quartile\ deviation = \frac{Q3 - Q1}{2}$$

The relative measure of quartile deviation is known as the coefficient of Q.D.

$$Coefficient\ of\ Q.D. = \frac{\frac{Q3 - Q1}{2}}{\frac{Q3 + Q1}{2}} = \frac{Q3 - Q1}{Q3 + Q1}$$

The larger the semi – interquartile range, the larger the spread of the central half of the data. Thus the semi –interquartile rang provides a measure of spread. Thus it indicate how closely the data are clustered around the median.

Mean deviation

Mean deviation is the average of the absolute values of the deviation scores; that is, mean deviation is the average distance between the mean and the data points. It is calculated as

$$\sum \frac{|\overline{X} - X_i|}{n}$$

Closely related to the measure of mean deviation is the measure of *variance*.

Variance

The variance is the most commonly accepted measure of variation. It represents the average of the squared deviations about the mean. Variance also indicates a relationship between the mean of a distribution and the data points; it is determined by averaging the sum of the squared deviations. Squaring the differences instead of taking the absolute values allows for greater flexibility in calculating further algebraic manipulations of the data. It is the average of the squared deviations between the individual scores and the mean. The larger the variance the more variability there is among the scores. When comparing two samples with the same unit of measurement (age), the variances are comparable even though the sample sizes may be different. Generally, however, smaller samples have greater variability among the scores than larger samples.

The average deviation from the mean is:

$$Ave.\ Dev = \frac{\sum(x - \mu)}{N}$$

The problem is that this summation is always zero. So, the average deviation will always be zero. That is why the average deviation is never used. So, to keep it from being zero, the deviation from the mean is squared and called the "squared deviation from the mean". This "average squared deviation from the mean" is called the variance. The formula for variance depends on whether you are working with a population or sample:

The formula for the variance in a population is where $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$ where μ is the mean and N is the number of scores.

When the variance is computed in a sample, the statistic $s^2 = \frac{\sum(X-M)^2}{N-1}$

where M is the mean of the sample and gives an unbiased estimate of σ^2.

Standard deviation

Standard deviation is the most familiar, important and widely used measure of variation. It is a significant measure for making comparison of variability between two or more sets of data in terms of their distance from the mean.

The standard deviation is the square root of the variance. It is denoted by σ and is computed as

$$\sigma = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n}}$$

The standard deviation has proven to be an extremely useful measure of spread in part because it is mathematically tractable. Many formulas in inferential statistics use the standard deviation. It possess the majority of the properties which are desirable in a measure of dispersion and is based on all observations. Because of these merits SD should always be used as the measure of dispersion unless there is some definite reason for selecting any other measure of dispersion.

Coefficient of Variation

The coefficient of variation is the ratio of the sample standard deviation to the sample mean. It is calculated as

$$Coefficient\ of\ variation\ (C.V.) = \frac{\sigma}{\bar{x}} * 100$$

It expresses the standard deviation as a percentage of the mean, so it can be used to compare the variability of two or more distributions even when the observations are expressed in different units of measurement. The coefficient of variation is a dimensionless number. So when comparing between data sets with different units or widely different means, one should use the coefficient of variation for comparison instead of the standard deviation. A standard application of the Coefficient of Variation is to characterize the variability of geographic variables over space or time. Coefficient of Variation is particularly applied to characterize the interannual variability of climate variables or biophysical variables. When coefficient of variable is lesser in the data, it is said to be more consistent or have less variability. On the other hand, the series having higher coefficient of variable has higher degree of variability or lesser consistency. When the mean value is close to zero, the coefficient of variation will approach infinity and is hence sensitive to small changes in the mean. Unlike the standard deviation, it cannot be used to construct confidence intervals for the mean.

Correlation

Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. The correlation analysis enables us to have an idea about the degree & direction of the relationship between the two variables under study. It is used to assess the possible linear association between two variables. If there is any relation between two variables *i.e.,* when one variable changes the other also changes in the same or in the opposite direction, we say that the two variables are correlated. Thus correlation is the study of existence, magnitude and direction of the relation between two or more variables. The measure of correlation called the correlation coefficient. If the ratio of change between two variables is uniform, then the correlation is said to be linear. If the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable, then the correlation is said to be non-linear or curvilinear. The nature of the graph gives us the idea of the linear type of correlation between two variables. If the graph is in a straight line, the correlation is called a "linear correlation" and if the graph is not in a straight line, the correlation is non-linear or curvi-linear.

Positive and negative correlation

If two variables change in the same direction *i.e.,* if one increases the other also increases, or if one decreases, the other also decreases), then this is called a positive correlation. If two variables change in the opposite direction *i.e.,* if one increases, the other decreases and vice versa), then the correlation is called a negative correlation. Through the coefficient of correlation, we can measure the degree or extent of the correlation between two variables. On the basis of the coefficient of correlation we can also determine whether the correlation is positive or negative and also its degree or extent.

If two variables changes in the same direction and in the same proportion, the correlation between the two is **perfect positive**. According to Karl Pearson the coefficient of correlation in this case is +1. On the other hand if the variables change in the opposite direction and in the same proportion, the correlation is **perfect negative** and its coefficient of correlation is -1. In practice we rarely come across these types of correlations.

If two variables exhibit no relations between them or change in variable does not lead to a change in the other variable, then we can say that there is **no correlation** between the two variables. In such a case the coefficient of correlation is 0.

Methods of Determining Correlation

The following are the most commonly used methods of determining correlation..

(1) Scatter Plot

(2) Karl Pearson's coefficient of correlation

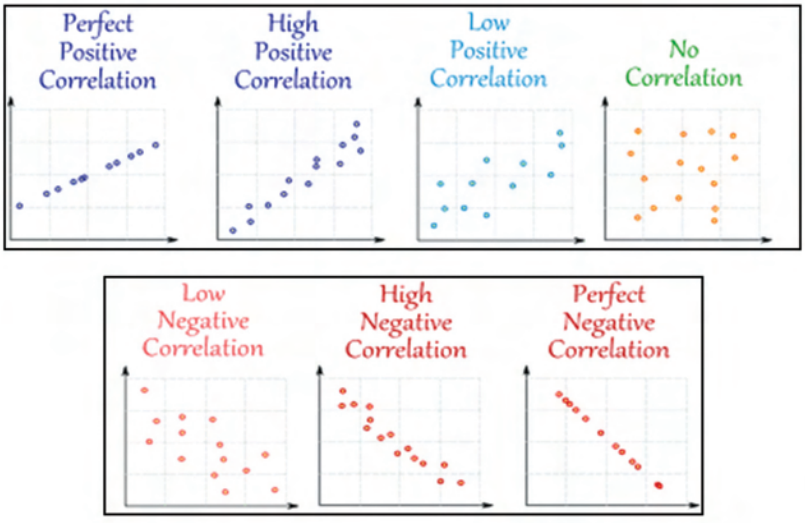
Scatter Plot (Scatter diagram or dot diagram)

The scatter diagram may be described as the diagram which helps us to visualize the relationship between two phenomena. This is the simplest method for finding out whether there is any relationship present between two variables. In this method the values of the two variables are plotted on a graph paper. One is taken along the x-axis and the other along the y-axis. By plotting the data, we get points on the graph which are generally scattered and hence the name 'Scatter Plot'. The manner in which these points are scattered, suggest the degree and the direction of correlation. The grater the scatter of the points on the chart, the lesser is the relationship between the two variables. The more closely the points come to a straight line, the higher the degree of relationship. The degree of correlation is denoted by 'r' and its direction is given by the signs positive and negative. Scatter diagrams will generally show one of five possible correlations between the variables:

- *Strong Positive Correlation :*The value of Y clearly increases as the value of X increases.
- *Strong Negative Correlation:* The value of Y clearly decreases as the value of X increases.
- *Weak Positive Correlation :* The value of Y increases slightly as the value of X increases.
- *Weak Negative Correlation:* The value of Y decreases slightly as the value of X increases.
- *No Correlation:* There is no demonstrated connection between the two variables.

Though this method is simple and provide a rough idea about the existence and the degree of correlation, it is not reliable. As it is not a mathematical method, it cannot measure the degree of correlation.

Illustrations



Karl Pearson's coefficient of correlation:

The most widely-used type of correlation coefficient is *Pearson r*, also called *linear* or *product- moment* correlation. It gives the numerical expression for the measure of correlation. The value of ' r ' gives the magnitude of correlation and sign denotes its direction. It is defined as

$$r = \frac{\sum XY}{n\sigma_x\sigma_y}$$

Where $X = (X_i - \bar{X})$, $Y = (Y_i - \bar{Y})$, $\sigma_x = s.d.of\ X$, $\sigma_y = s.d.of\ Y$ and n is the number of pairs of observations

Properties of Correlation coefficient

- The value of correlation does not depend on the specific measurement units used; for example, the correlation between height and weight will be identical regardless of whether *inches* and *pounds*, or *centimeters* and *kilograms* are used as measurement units.
- The value of correlation coefficient lies between -1 and +1, -1 means perfect negative linear correlation and +1 means perfect positive linear correlation.
- The correlation coefficient r only measures the strength of a linear relationship. There are other kinds of relationships besides linear.
- If the two variables are independent, then the value of the correlation coefficient is zero. If the value of the correlation coefficient is zero, it does not mean that there is no correlation, but there may be non-linear correlation.
- The value of r does not change if the independent (x) and dependent (y) variables are interchanged.
- The correlation coefficient r does not change if the scale on either variable is changed. You may multiply, divide, add, or subtract a value to/from all the x-values or y-values without changing the value of r.
- The correlation coefficient r has a Student's t distribution.

Assumptions to use the Pearson product-moment correlation

- The measures are approximately normally distributed
- The variance of the two measures is similar (homoscedasticity)

- The relationship is linear
- The sample represents the population
- The variables are measured on a interval or ratio scale

Testing the Significance of the Correlation Coefficient

The correlation coefficient, *r*, tells us about the strength and direction of the linear relationship between *x* and *y*. However, the reliability of the linear model also depends on how many observed data points are in the sample. We need to look at both the value of the correlation coefficient *r* and the sample size *n*, together.

We perform a hypothesis test of the "significance of the correlation coefficient" to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population. The sample data are used to compute *r*, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only have sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, *r*, is our estimate of the unknown population correlation coefficient. The hypothesis test lets us decide whether the value of the population correlation coefficient σ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient *r* and the sample size *n*.

The correlation coefficient r has a t distribution with n-2 degrees of freedom. The test statistic used is

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is significant and there exists a linear relationship between the two variables. If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is not significant and there is insufficient evidence to conclude that there is a significant linear relationship between the two variables.

Regression analysis

Regression analysis is a statistical tool used for the investigation of relationships between variables. It is the study of *linear*, *additive* relationships between variables. Correlation gives us a measure of the magnitude and direction between variables. It is a technique used for predicting the unknown value of a variable from the known value of another variable. When

there is only one independent variable then the relationship is expressed by a straight line. This procedure is called simple linear regression or bivariate regression. More precisely, if X and Y are two related variables, then linear regression analysis helps us to predict the value of Y for a given value of X. Multiple regression is an extension of bivariate regression in which several independent variables are combined to predict the dependent variable. In multiple regression analysis, the value of Y is predicted for given values of X_1, X_2, ..., X_k. This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

Dependent and Independent Variables

By simple linear regression, we mean models with just one independent and one dependent variable. The variable whose value is to be predicted is known as the dependent variable and the one whose known value is used for prediction is known as the independent variable. Similarly for Multiple Regression the variable whose value is to be predicted is known as the dependent variable and the ones whose known values are used for prediction are known independent variables.

The Regression Model

The line of regression of Y on X is given by Y = a + bX where a and b are unknown constants known as intercept and slope of the equation. This is used to predict the unknown value of variable Y when value of variable X is known.

The Simple Linear Regression model is

$$Y = a + bX$$

The **Regression Coefficient** is the constant 'b' in the regression equation that tells about the change in the value of dependent variable X corresponding to the unit change in the independent variable Y and can be represented as:

$$b = r\frac{\sigma_x}{\sigma_y}$$

Where r is the correlation coefficient σ_x is the standard deviation of x, σ_y is the standard deviation of y.

In general, the multiple regression equation of Y on X_1, X_2, ..., X_k is given by:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots\ldots\ldots + b_k X_k$$

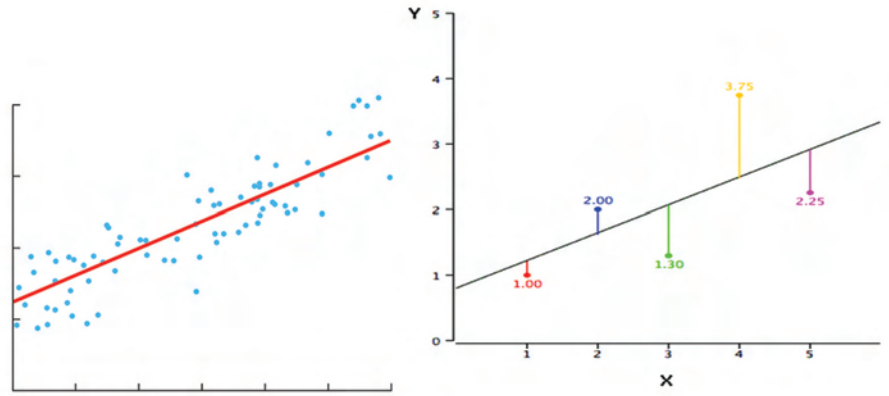
Here b_0 is the intercept and b_1, b_2, b_3, ..., b_k are analogous to the slope in linear regression equation and are also called regression coefficients. They can be interpreted as

the change in the value of dependent variable (Y) corresponding to unit change in the value of independent variable X_i.

Fitting of regression line

In scatter plot, we have seen that if the variables are highly correlated then the points (dots) lie in a narrow strip. If the strip is nearly straight, we can draw a straight line, such that all points are close to it from both sides. Such a line can be taken as an ideal representation of variation. This line is called the line of best fit if it minimizes the distances of all data points from it and also called as the line of regression. Now prediction is easy because all we need to do is to extend the line and read the value. Thus to obtain a line of regression, we need to have a line of best fit.

The problem of choosing the best straight line then comes down to finding the best values of a and b. By 'best' we mean the values of a and b that produce a line closest to all n observations. This means that we find the line that minimizes the distances of each observation to the line. Choose the values of a and b that give the line such that the sum of squared deviations from the line is minimized. This method of estimation of parameters is called least square method. The best line is called the regression line, and the equation describing it is called the regression equation. The deviations from the line are also called residuals.



R^2 - coefficient of determination

Once a line of regression has been constructed, one can check how good it is (in terms of predictive ability) by examining the coefficient of determination (R^2), which is defined as the proportion of variance of the dependent variable that can be explained by the independent variables. The coefficient of determination is a measure of how well the regression equation y = a + bX performs as a predictor of y. Its value represents the percentage

of variation that can be explained by the regression equation. R^2 always lies between 0 and 1. Higher values of this are generally taken to indicate a better model. A value of 1 means every point on the regression line fits the data; a value of 0.5 means only half of the variation is explained by the regression. The coefficient of determination is also commonly used to show how accurately a regression model can predict future outcomes.