www.prepnext.com

Notes Prepared By:
RAKESH AGARWAL
M.Com, MBA, FIII
E-mail: messagerakesh@gmail.com
WhatsApp No: 8486118428

Business Statistics

Unit 2

Q: What do you mean by correlation? Mention the uses of correlation (co-variation). *(www.prepnext.com)*

Or

Explain the utility or significance of correlation.

Ans.:

Correlation is the relationship that exists between two or more variables. If two variables are related to each other in such a way that the change in one creates a corresponding change in the other than the variables are said to be correlated.

According to **Prof. Ya-Lun-Chou,** "Correlation analysis attempts to determine the 'degree of relationship' between variables."

According to **Prof. Simpson and Kafka,** "Correlation analysis deals with the association between two or more variables."

According to **A.M. Tuttle,** "Correlation is an analysis of the covariation between two or more variables."

According to **L.R. Conner**, "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the other(s) then they are said to be correlated."

From the above definitions we can say that the correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables. For example:- relationship between the heights and weights, relationship between the price of commodity and demand of commodity.

Uses (Utility) of correlation

The important uses of correlation are:

- **Degree of relationship**: With the help of correlation analysis we can measure in one figure the degree of relationship existing between the variables. For example:- relationship between price and supply, income and expenditure, etc.
- **Helps in economic theory and business studies:** Correlation analysis helps in deriving the relationship between variables in economic theory. It helps in understanding the economic behavior, and in locating the critically important variables on which others depend
- **Reduces the range of uncertainty**: the effect of correlation is to reduce the uncertainty of our prediction.
- Correlation analysis **contributes to the understanding of economic behavior**.

CORRELATION AND CAUSATION

Correlation analysis helps us in determining the degree of relationship between two or more variables – it does not tell us anything about cause and effect relationship. The explanation of a significant degree of correlation may be any one, or combination of the following reasons:

(i) **The correlation may be due to pure chance, especially in a small sample**: Such a correlation may arise either because of pure random sampling variation or because of the bias of the investigator in selecting the sample.

(ii) **Both the correlated variables may be influenced by one or more other variables**: It is just possible that a high degree of correlation between the variables may be due to some causes affecting each variable or different causes affecting each with the same effect. For example, a high degree of correlation between the yield per acre of rice and tea may be due to the fact that both are related to the amount of rainfall. But none of the two variables is the cause of the other.

(iii) **Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect**: There may be high degree of correlation between the variables but it may be difficult to pinpoint as to which is the cause and which is the effect. For example, such variables as demand and supply, price and production, etc. mutually interact. At times it may become difficult to explain from the two correlated variables which is the cause and which is the effect because both may be resting on each other.

Q: What are the merits and demerits of correlation?

(www.prepnext.com)

Ans.:

MERITS:

1) Correlation provides the direction and degree of relationship between two variables.

2) The correlation gives precise measurement of relationship between two variables that can be meaningfully interpreted and made use of

3) The coefficient of correlation is independent of the changes in the scale and origin of the two variables

DEMERITS:

1) The extreme value of an item unduly affects correlation coefficient

2) The limits of coefficient correlation (between +1 to –1) needs to be carefully understood and interpreted.

3) Correlation takes into account linear relationship only.

Q: Explain the different types of correlation? *(www.prepnext.com)*

Ans.:
Correlation is classified in several different ways. Three of the most important ways of classifying correlation are:

1. Positive or Negative
2. Simple, partial and multiple
3. Linear and non-linear

Positive and Negative Correlation:

Whether correlation is positive (direct) or negative (inverse) would depend upon the direction of change of the variables.

1. **Positive correlation:-** If both the variables vary in the same direction, correlation is said to be positive. In other words, if one variable increases, the other also increases or if one variable decreases the other also decreases.
2. **Negative correlation:-** If both the variables vary in opposite direction, the correlation is said to be negative. In other words, if one variable increases but the other variable decreases or if one variable decreases but the other variable increases, then the correlation between the two variables is said to be negative. For example:- correlation between price and demand.

Simple, Partial and Multiple Correlation:

The distinction between simple, partial and multiple correlation is based upon the number of variables studied.

1. **Simple correlation:-** When only two variables are studied it is a case of simple correlation. For example:- when we study relationship between production of wheat and the amount of rainfall, it is a case of simple correlation.
2. **Multiple correlation:-** When three or more variables are studied simultaneously it is a case of multiple correlation. For example- when we study the relationship between production of wheat, amount of rainfall and amount of fertilizers used, it is a case of multiple correlation.
3. **Partial Correlation:** In partial correlation we recognize more than two variables, but consider only two variables to be influencing each other, the effect of other influencing variables kept constant. For example, when we study the relationship between the production of rice, amount of rainfall and amount of fertilizers used, but limit our correlation analysis of production and fertilizers to periods when a certain average daily temperature existed it becomes a problem relating to partial correlation only.

Linear and Non-Linear (Curvilinear) Correlation:

The distinction between linear and non-linear correlation is based upon the constancy of the ratio of change between the variable.

1. **Linear correlation:-** If the amount of change in one variable bears a constant ratio to the amount of change in other variable then the correlation is said to be linear. If such variables are plotted on a graph paper all the plotted points would fall on a straight line.
2. **Non-linear correlation (curvilinear):-** 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. For example, if we double the amount of rainfall the production of rice or wheat, etc., would not necessarily be doubled. If such variables are plotted on a graph paper, the graph will be a curve.

Q: What do you mean by co-efficient of correlation? Mention its properties. *(www.prepnext.com)*

Ans.:

The degree to which the two variables are inter-related is measured by coefficient which is called the coefficient of correlation. It gives the degree of correlation. The coefficient of correlation between the two variables x and y are generally denoted by r or r_{xy}

Properties:-

1. Coefficient of correlation is a measure of the closeness of a fit in a relative sense.
2. Correlation coefficient lies between -1 and +1
3. The correlation is perfect and positive if r = 1 and it is perfect and negative if r = – 1
4. If r = 0 then, there is no correlation between the two variables and thus, the variables are said to be independent.

5. The correlation coefficient is a pure number and is not affected by a change of origin and scale in magnitude.
6. The coefficient of correlation is the geometric mean of two regression coefficient.

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

Q: What are the different methods of studying correlation?

(www.prepnext.com)

Ans.:

FEW METHODS OF STUDYING CORRELATION

The various methods of ascertaining whether two variables are correlated or not are:

1. Scatter Diagram Method
2. Graphic Method
3. Karl Pearson's Coefficient of Correlation
4. Spearman's Correlation Coefficient

SCATTER DIAGRAM METHOD:

The simplest device for ascertaining whether two variables are related is to prepare a dot chart called scatter diagram. When this method is used the given data are plotted on a graph paper or simply scatter plot in the form of dots, i.e., for each pair of X and Y values we put a dot and thus obtain as many point as the number of observations. By looking to the scatter of the various points we can form an idea as to whether the variables are related or not. The greater the scatter of the plotted points on the chart, the lesser is the relationship between the two variables.

If the plotted points fall in a narrow band there would be a high degree of correlation between the variables. Correlation shall be positive, if the points show a rising tendency from the lower left-hand corner to the upper right-hand corner.

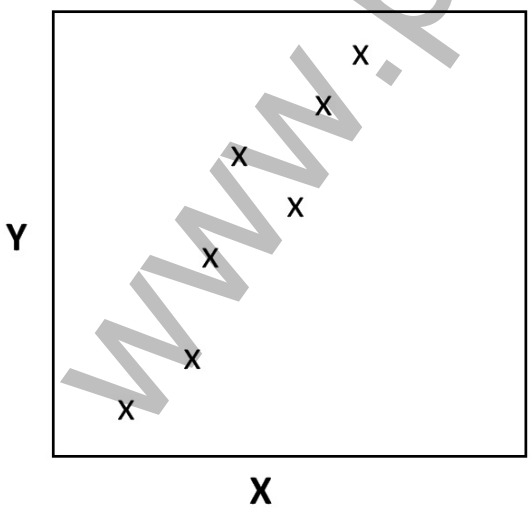
Correlation shall be negative if the points show a declining tendency from the upper left-hand corner to the lower right-hand corner of the diagram.

If all the points are lying on a straight line falling from the upper left-hand corner to the lower right-hand corner of the diagram, correlation is said to be perfectly negative (i.e., r = − 1).

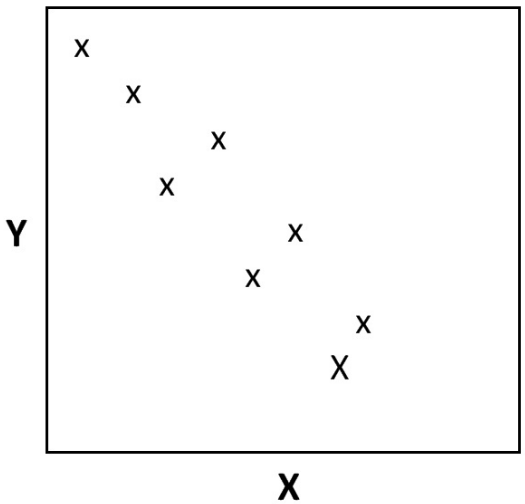
If all the points are lying on a straight line rising from the lower left-hand corner to the upper right-hand corner of the diagram, correlation is said to be perfectly positive (i.e., r = + 1).

If the plotted points lie on a straight line parallel to the X-axis or in a haphazard manner, it shows absence of any relationship between the variables (i.e., r = 0).

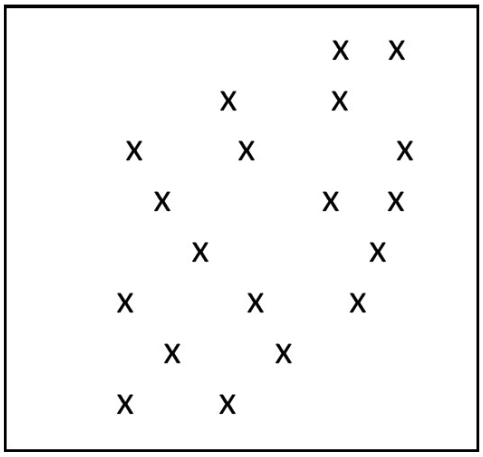
High Degree of Positive Correlation



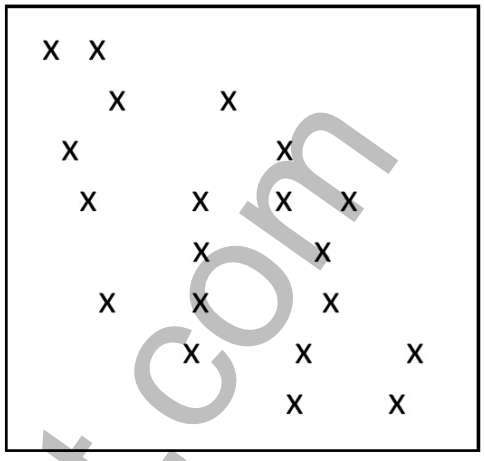
High Degree of Negative Correlation



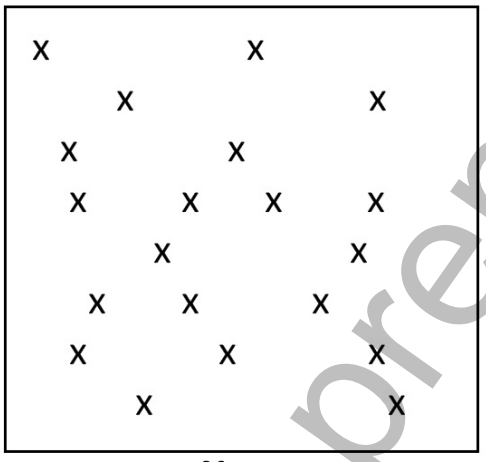
Low Degree of Positive Correlation



Low Degree of Negative Correlation



No Correlation (r = 0)



MERITS AND LIMITATIONS OF SCATTER DIAGRAM METHOD

Merits:

1. It is simple and non-mathematical method of studying correlation between the variables. As such it can be easily understood and a rough idea can very quickly be formed as to whether or not the variables are related.
2. It is not influenced by the size of extreme items.

Limitations:

This method gives an idea about the direction of correlation and also whether it is high or low. But we cannot establish the exact degree of correlation between the variables as is possible by applying the mathematical methods.

GRAPHIC METHOD:

When this method is used the individual values of the two variables are plotted on the graph paper. We thus obtain two curves, one for X variable and another for Y variables. By examining the direction and closeness of the two curves so drawn we can infer whether or not the variables are related. If both the curves drawn on the graph are moving in the same direction (either upward or downward) correlation is said to be positive. On the other hand, if the curves are moving in the opposite directions correlation is said to be negative.
This method is normally used where we are given data over a period of time. i.e., in case of time series. However, in this method we cannot get a numerical value describing the extent to which the variables are related.

KARL PEARSON'S COEFFICIENT OF CORRELATION:

Of the several mathematical methods of measuring correlation, the Karl Pearson's method, popularly known as Pearson's coefficient of correlation is most widely used in practice. The Pearson's coefficient of correlation is denoted by the symbol r. The formula for computing Pearsonian r is:

$$r = \sum xy \;/\; (N \partial_x \partial_y)$$

Here, $x = (X - \bar{X})$; $y = (y - \bar{y})$;

∂_x = Standard deviation of series X;

∂_y = Standard deviation of series Y

N = Number of pairs of observations

r = the (product moment) correlation coefficient

Assumption of the Pearsonian Coefficient:

Karl Pearson's coefficient of correlation is based on the following assumptions:

1. There is linear relationship between the variables i.e., when the two variables are plotted on a scatter diagram a straight line will be formed by the points so plotted.
2. *The two variables under study are affected by a large number of independent causes so as to form a normal distribution. Variables like height, weight, price, demand, supply, etc. are affected by such forces that a normal distribution is formed.*
3. There is a cause and effect relationship **between the forces affecting the distribution** of the items in the two series. If such a relationship is not formed between the variables, i.e., if the variables are independent, there cannot be any correlation. For example, there is no relationship between income and height because the forces that affect these variables are not common.

MERITS AND LIMITATIONS:

MERITS:

Karl Pearson's method is the most popular mathematical method used for measuring the degree of relationship. The correlation coefficient summarises in one figure not only the degree of correlation but also the direction, i.e., whether correlation is positive or negative.

Limitations:

The chief limitations of the method are:

1. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is correct or not.
2. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

3. The value of the coefficient is unduly affected by the extreme items.
4. As compared with other methods this method takes more time to compute the value of correlation coefficient.

INTERPRETING COEFFICIENT OF CORRELATION:

The coefficient of correlation measures the degree of relationship between two sets of figures. The following general rules are given which would help in interpreting the value of r:

1. When r = + 1, it means there is perfect positive relationship between the variables.
2. When r = – 1 , it means there is perfect negative relationship between the variables.
3. When r = 0, it means there is no relationship between the variables, i.e., the variables are uncorrelated.
4. The closer r is to + 1 or – 1, the closer the relationship between the variables and the closer r is to 0, the less close the relationship.

PROPERTIES OF THE COEFFICIENT OF CORRELATION:

1. The coefficient of correlation lies between –1 and +1. Symbolically, $-1 \leq r \leq +1$.
2. The coefficient of correlation is independent of change of scale and origin of the variable X and Y.
3. The coefficient of correlation is the geometric mean of two regression coefficients.
   Symbolically,

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

4. The degree of relationship between the two variables is symmetric as shown below:

$$r_{xy} = r_{yx}$$

$$r_{xy} = \sum xy / (N \partial_x \partial_y) = \sum yx / (N \partial_y \partial_x) = r_{yx}$$

RANK CORRELATION COEFFICIENT:

This method of finding out co-variability or the lack of it between the two variables was developed by Charles Edward Spearman in 1904. This measure is especially useful when quantitative measures for certain factors (such as in the evaluation of leadership ability or the judgment of female beauty) cannot be fixed, but the individual in the group can be arranged in order thereby obtaining for each individual a number indicating his (her) rank in the group. Spearman's rank correlation coefficient is defined as:

$$R = 1 - [6\sum d^2 \ / \ (N^3 - N)]$$

Where, R denotes rank coefficient of correlation and D refers to the difference of rank between paired items in two series.

FEATURES OF SPEARMAN'S CORRELATION COEFFICIENT:

1. The value of Karl Pearson's correlation coefficient ranges between + 1 and – 1.
2. The sum of the differences of ranks between two variables shall be zero. Symbolically, $\sum d = 0$.
3. Spearman's correlation coefficient is distribution-free or non-parametric *because no strict assumptions are made about the form of population from which sample observations are drawn.*
4. The Spearman's correlation coefficient is nothing but Karl Pearson's correlation coefficient between the ranks. Hence, it can be interpreted in the same manner as Pearson's correlation coefficient.

MERITS AND LIMITATIONS OF THE RANK METHOD:

Merits:

1. This method is simpler to understand and easier to apply compared to the Karl Pearson's method. The answers obtained by this method and the Karl Pearson's method will be the same provided no value is repeated, i.e., all the items are different.
2. Where the data are of a qualitative nature like honesty, efficiency, intelligence, etc., this method can be used with great advantage. For example, the workers of two factories can be ranked in order of efficiency and the degree of correlation established by applying this method.
3. This is the only method that can be used where we are given the ranks and not the actual data.
4. Even when actual data are given, rank method can be applied for ascertaining correlation.
5. Rank correlation is very useful when the data are non-normally distributed.

Limitations:

1. This method cannot be used for finding out correlation in a grouped frequency distribution.
2. Where the number of items exceeds 30 the calculations become quite tedious and require a lot of time. Therefore, this method should not be applied where N exceeds 30 unless we are given the ranks and not the actual values of the variables.

WHEN TO USE RANK CORRELATION COEFFICIENT?

The rank method has principal use:

1. If the initial data are in the form of ranks.
2. If N is fairly small (say, not more than 25 or 30) rank method is sometimes applied to interval data as an approximation to the more time-consuming 'r'. This requires that the interval data be transferred to rank orders for both variables. If N is much in excess of 30, the labour required in ranking the scores becomes greater than is justified by the anticipated saving of time through the rank formula.

Q: Define Regression? *(www.prepnext.com)*

Ans.:

Regression is the measure of average relationship between two or more variables in terms of the original units of the data. In other word, regression shows a relationship between the average values of two variables. Thus, regression is very helpful in estimating and predicting the average value of one variable for a given value of other variable.

1. According to **Morris Hamburg**, "The term 'regression analysis' refers to the methods by which estimates are made of the values of a variable from a knowledge of the values of one or more other variables and to the measurement of the errors involved in this estimation process."
2. According to **Ya-Lun Chou**, "Regression analysis attempts to establish the 'nature of relationship' between variables – that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting."
3. "Regression is the measure of the average relationship between two or more variables in terms of the original units of the data."

It is clear from the above definitions that regression analysis is a statistical device with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable. The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent variable or "explained" variable.

Q: What are the uses of regression analysis? *(www.prepnext.com)*

Ans.:

1. Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. The study of regression is of considerable help to the economists and businessmen. In economics it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory and economic life. For example, if we know that two variables, price (X) and demand (Y), are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X.
2. Regression analysis provides estimates of values of the dependent variable from values of the independent variable.
3. *Regression analysis also obtains a measure of the error involved in using the regression line as a basis for estimation. For this purpose the standard error of estimate is calculated.*
4. With the help of regression coefficients we can calculate the correlation coefficient.

Q: What are the different kinds of regression analysis?

(www.prepnext.com)

Ans.:

A) Simple Vs multiple regression:

In simple regression, there is one dependent variable and one independent variable. Here the dependent variable is estimated on the basis of another independent variable. But, in case of multiple regression, there are one dependent variables and two or more independent variables.

B) LINEAR VS CURVILINEAR REGRESSION:

In linear regression, the relationship between variable can be presented by a straight line. But, in curvilinear regression, there is no linear relationship between the variables. It deals with relationships like parabolic, hyperbolic, etc.

Q: What do you mean by lines of regression? Why there are two lines of regression? *(www.prepnext.com)*

Ans.:

A line of regression is the line which gives the best estimate of one variable X for any given value of the other variable Y.

There are two lines of regression because one equation is used for estimating the value of X variable for a given value of Y variable and the second equation is used for estimating the value of Y variable for a given value of X variables. In both cases, the assumption is that one is an independent variable and the other is a dependent variable and vice versa.

Q: Mention the properties of Regression Coefficient?

(www.prepnext.com)

Ans.:

There are two regression coefficient b_{yx} and b_{xy}

The regression coefficient of Y on X is $b_{yx} = r\ (\partial y/\partial x)$

The regression coefficient of X on Y is $b_{xy} = r\ (\partial x/\partial y)$

Properties:-

1. The correlation coefficient is the geometric mean of regression coefficient, that is

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

2. Since the value of the coefficients of correlation (r) cannot exceed one, one of the regression coefficients must be less than one or, in other words, both the regression coefficients cannot be greater than one. If one of the regression coefficient is greater than unity than the other is less than unity.
3. Arithmetic mean of the regression coefficient is greater than the correlation coefficient
4. Regression coefficient are independent of change of origin but not of scale
5. Both regression coefficients will have the same sign either both positive or both negative
6. The sign of correlation coefficient is the same as that of regression coefficients. i.e., if regression coefficients have a negative sign, r will also be negative; and if regression coefficients have a positive sign, r would also be positive.
7. Since $b_{xy} = r\ (\partial x/\partial y)$, we can find out any of the four values given the other three. For example, if we know that r = 0.6, ∂x = 4 and $b_{xy} = 0.8$, we can find ∂y.

$$b_{xy} = r\ (\partial x/\partial y)$$

Substituting the given values:

$0.8 = (0.6 \times 4)/\ \partial y$ Or $\partial y = 2.4/\ 0.8 = 3$

Q: What are the limitations of Regression Coefficient?

(www.prepnext.com)

Ans.:

LIMITATIONS OF REGRESSION ANALYSIS

1) In making estimate from a regression equation, it is assumed that relationship has not changed since the regression equation was computed.

2) The relationship shown by the regression equation may not be the same if the equation is extended beyond the values used in computing the equation. For example, there may be a close linear relationship between the yield of a crop and the amount of fertilizer applied, with the yield increasing as the amount of fertilizer is increased. It would not be logical, however, to extend this equation beyond the limits of the experiment for it is quite likely that if the amount of fertilizer were increased indefinitely, the yield would eventually decline as too much fertilizer was applied.

3) The relationship between two variables can be strongly influenced by other variables that are lurking in the background (a Lurking variable has important effect on the relationship among the variables in a study but is not included among its variables studied). Lurking variables which are often unrecognized and unmeasured can dramatically change the conclusions of a regression study.

Q: Differentiate between correlation and regression.

www.prepnext.com)

Ans.:

The main differences between correlation and regression are:-

CORRELATION	REGRESSION
Coefficient of correlation is a measure of degree of covariability between X and Y.	The objective of regression analysis is to study the 'nature of relationship' between the variables so that we may be able to predict the value of one on the basis of another.
In correlation analysis, r_{xy} and r_{yx} are symmetric ($r_{xy} = r_{yx}$), i.e., it is immaterial which of X and Y is dependent variable and which is independent variable.	In regression analysis the regression coefficients b_{xy} and b_{yx} are not symmetric, i.e., $b_{xy} \neq b_{yx}$ and hence it definitely makes a difference as to which variable is dependent and which is independent
It measures degree and direction of relationship between the variables	It measures the nature and extent of average relationship between 2 or more variables in terms of the original units of the data
It is a relative measure showing association between variables	It is an absolute measure of relationship
Correlation coefficient is independent of change of both origin and scale	Regression coefficient is independent of change of origin but not of scale
Correlation coefficient is independent of units of measurement	Regression coefficient is not independent of units of measurement
There may be nonsense correlation between two variables which is purely due to chance and has no practical relevance such as increase in weight and increase in income of a group of people.	There is nothing like nonsense regression

It is not a forecasting device	It is a forecasting device which can be used to predict the value of dependent variable from the given value of independent variable
Does not study cause-effect relationship	Regression analysis attempts to find the cause-effect relationship between the variables
Expression of a relationship between the variable ranges from -1 to +1	Regression coefficients do not have any such boundaries.

N.B.

1. **When there is either perfect positive or perfect negative correlation between the two variables (r = ± 1) the regression lines will coincide, i.e, we will have only one line**. The farther the two regression lines from each other, the lesser is the degree of correlation and the nearer the two regression lines to each other, the higher is the degree of correlation. If the variables are independent, r is zero and the lines of regression are at right angles, i.e., parallel to OX and OY.
2. The regression lines cut each other at the point of average of X and Y, i.e., if from the point where both the regression lines cut each other a perpendicular is drawn on the X-axis, we will get the mean value of X and if from that point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.
3. Regression lines are drawn on least squares assumption which stipulates that the sum of squares of the deviations of the observed 'Y' values from the fitted line shall be minimum. The total of the squares of the deviations of the various points is minimum only from the line of best fit.