1. Introduction

Welcome to the series of E-learning modules on Concept of error in Regression. In this module, we are going to study about the regression, its comparison with the correlation, regression analysis and its importance, methods of studying regression, error in regression and the root mean square error.

By the end of this session, you will be able to:

- Explain the meaning of regression
- Explain the application of regression
- Describe the comparison of regression with correlation
- Explain the importance or utility of regression analysis
- Explain the methods of studying regression
- Describe the error in Regression
- Explain the root mean square error

Before we discuss about the error in regression, let us discuss about the regression.

After studied the fact that the two variables are closely related we may be interested in estimating or predicting the value of the one variable given with the value of another.

For example, if we know that advertising and sales are correlated we may find out the expectation amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

Similarly, if we know that the yield of rice and rainfall are closely related we may find out the amount of rain required to achieve a certain production figure.

The statistical tool with the help of which we are in a position to estimate or predict the unknown vales of one variable from known values of another variable is called regression. With the help of regression analysis, we are in a position to find out the average probable change in one variable given with a certain amount of change in another.

According to W. Z. Harisch, “Correlation analysis tests the closeness with which two or more phenomena co-vary; regression analysis measures the nature and extent of the relation, thus enabling us to make predictions.”

Now, let us see what the Meaning of regression is.
The dictionary meaning of the term regression is the act of returning or going back.

The term regression was first used by Francis Galton towards the end of nineteenth century while studying the relationship between the height of fathers and sons.

This term was introduced by him in the paper ‘Regression towards Mediocrity in Hereditary Stature’. His study of height of about the thousand father and sons revealed a very interesting relationship, that is tall fathers tend to have tall sons and short fathers tend to have short sons, but the average height of the sons of a group of tall fathers less than that of the fathers

and average height of the sons of a group of short fathers is greater than that of the fathers. The line describing the tendency to regress or going back was called by Galton a ‘Regression Line’.

The term is still used to describe that, line drawn for a group of points to represent the trend present, but it no longer necessarily carries the original implication that Galton intended.

These days there is growing tendency of the modern writers to use the term estimating line instead of regression line because the expression estimating line is more clarificatory in character.

Regression analysis is a branch of statistical theory that is widely used in almost all the scientific disciplines. In economics, it is the basic technique for measuring or estimating the relationship among economic variables that constitute the essence of economic theory of an 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.

Thus, we find that the study of regression is of considerable help to the economists and businessmen. The uses of regression are not confined to the economics and business fields only, but its implications are extended to almost all the natural, physical and social sciences. Regression analysis is one of the very scientific techniques for making such predictions.

According to M. M. Blair, “Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data”.

It is mentioned earlier that regression is used when there is a cause and effect relation between the two variables. We name the two variables as dependent and independent variable.

The variable whose value is influenced or is to be predicted is called dependent variables, which is regressed or explained variable. The variable, which influences the values or used for prediction, is called independent variable, regressor, predictor, or explanatory variable.

2. Utility of Regression Analysis

Now, let us see the utility of regression analysis.
By above definition it is clear that regression analysis is done for estimating or predicting the unknown value of one variable form the known value of the other variable. This is very useful statistical tool, which is used in both natural and social sciences.

In the field of business, this tool of statistical analysis is very widely used. Businesspersons are interested in predicting future production, consumption, investment, prices, profits, sales etc. In fact, the success of a businessman depends on the correctness of the various estimates that he is required to make.

In sociological studies and in the field of economic planning, projections of population, birth rates, death rates and other similar variables are of great use.
In our day-to-day life, we come across many variables, which are inter-related.

For example, with a rise in price, the demand of a commodity goes down.

Or with better monsoons the output of agricultural product increase or the effect of expenditure on publicity may lead to rise in the volume of sales. With the help of regression analysis, we can estimate or predict the effect of one variable on the other.

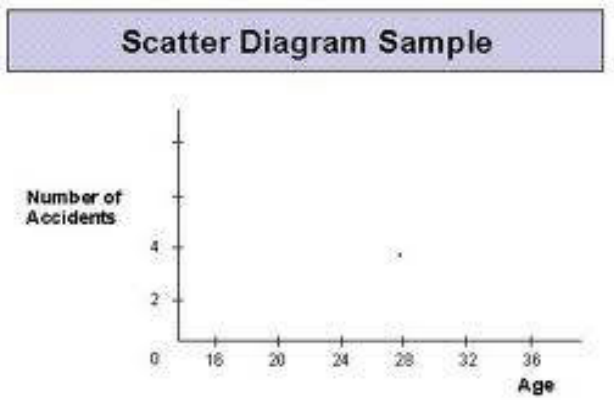
For example, we can also predict the fall in demand when there is a price rises by a particular amount. However, in social sciences there is multiple causation, which means that a large number of factors affect various variables. The regression study, which confines itself to a study of only two variables, is called simple regression. The regression analysis, which studies more than two variables at a time, is called multiple regressions.

3. Various Methods of Studying Regression

Now let us discuss the various methods of studying regression.
Broadly speaking, regression can be studied either graphically or algebraically.

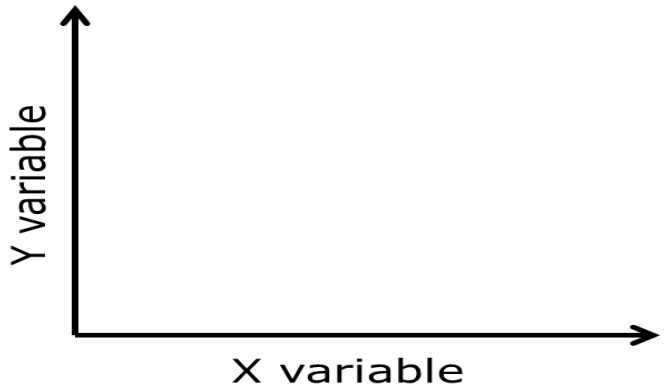
- Graphical study of regression: When regression is studied with the help of graphic methods, we have to draw a scatter diagram. A scatter diagram contains one point for one pair of values of X and Y variable.

Figure 1



Usually X variable is shown on the horizontal scale and Y variable on the vertical scale. When all related pair of values have been plotted on a scatter diagram, we have to draw 2 regression lines to predict the values of X and Y variables.

Figure 2



The regression line, which is used to predict the values of Y for a value of X is called the Regression line of Y on X. Similarly, the regression line, which is used to predict a value of X

for a value of Y, is called regression line of X on Y.

If the coefficient of correlation between X and Y is perfect, that is, its value is either plus 1 or minus 1, then there will be only one regression line, as the variations in the two series in such cases always increase or decrease by a constant figure.
In other words, we say that the two regression lines will be identical if the correlation between the two variables is perfect.

- Algebraic method: Here we use method of least squares to fit the regression lines by predicting the type of relation between the two variables.

Let us discuss the Difference between correlation and Regression.

Correlation measures the relationship between two variables, which vary in the same or opposite directions.
Whereas, regression means going back or act of return. It is a mathematical measure, which shows the average relationship between two variables.

In correlation, both X and Y variables are random variables.
Whereas, in regression X is random variable and Y is a fixed variable. However, both variables may be random variables.

Correlation tells the degree of relationship between the two variables and not the cause and effect of the variables.
However, Regression points out the cause and effect of relationship between the variables. It establishes a functional relationship.

Correlation is confined only to linear relationship between two variables.
Regression studies both linear and non-linear relationship between the variables.

Correlation is immaterial whether X variable depends on Y and Y depends on X variable.
In regression, there is a functional relationship between the two variables.

There can be nonsense or spurious correlation between the two variables.
There is no such nonsense regression equation.

Coefficient of correlation is not affected by change of scale or change of origin.
Regression coefficient is independent of change of origin by not of change of scale.

In coefficient, both the variables must be positively associated or vice-versa.
Regression explains that the decrease in one variable increase in the other variable.

4. Importance of Regression Analysis

Now, let us discuss the Importance of regression analysis.
By the study of regression analysis, we are able to obtain the most probable values of one variable from the known values of another variable.

For example, if two series relating to price and supply are correlated, we can find out what would be the effect on price if the supply of the commodity has increased or decreased to a particular level. The regression can also be used in natural, physical, and social science.

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

According to Taro Yamane, "One of the most frequently used technique in economics and business research to find a relation between two or more variables that are related casually, is regression analysis".

Now, let us discuss about the Concept of Error in Regression.

The theory formulates exact functional relationship among the variables. However, dealing with common data even an ordinary investigator will feel that not all observations fall exactly on a straight line or on any other smooth functions.

The best we can expect is that the observed quantities will be closer to the line, that is, why our regression model requires extension and stochastic disturbance term. The introduced term is known as disturbance or error term, because it represents the effects of all those factors, which are not suspected by the investigator.

Error term may have positive or negative values and is drawn at random. Measurement also causes error in the model. If investigator has collected wrong data, the observed outcome will contain two ingredients, the theoretical prediction and experimental error. We assume that the error term is distributed normally with mean zero and variance sigma square.

We make three assumptions about the regularity of these distributions.

1. Homogeneous variance: The distribution of the dependent variable has the same spread. Formally, this means that the probability distribution has the same variance sigma square.
2. Linearity: The mean of the dependent variable is a straight line, known as the true population regression line.
3. Independence: The random variables are statistically independent.

The error exists because of:

1. Measurement error: There are various reasons why the dependent variable is not measured correctly. For example, in a study of consumption of families at various

income levels, the measurement error in consumption might consist of

2. Budget and reporting inaccuracies.
3. Inherent Variability: Even if there were no measurement error, repetition of an experiment using exactly same amount of income would result in different levels of consumption.

The regression line generally does not go through all the data: approximating the data using the regression line entails some error.

The vertical amount by which the line misses a datum is called a RESIDUAL—it is the error in estimating the value of Y for that datum from its value of X using the regression line.

The root mean square (RMS) of the residuals has a simple relation to the correlation coefficient and the standard deviation of Y: It is
1 minus r square power half into Standard deviation of Y.

The regression line does not pass through all the points on the scatter plot exactly unless the correlation coefficient is plus or minus 1. In general, the data are scattered around the regression line. Each datum will have a vertical residual from the regression line; the sizes of the vertical residuals will vary from datum to datum. The root mean square of the vertical residuals measures the typical vertical distance of a datum from the regression line.

5. Root Mean Square

The root mean square is a measure of the typical size of the elements in a list. Thus, the root mean square of the vertical residuals is a measure of the typical vertical distance from the data to the regression line, which is the typical error in estimating the value of Y by the height of the regression line. A bit of algebra shows that the root mean square of the vertical residuals from the regression line or the root mean square error of regression is:
1 minus r square power half into Standard deviation of Y.

The root mean square error of regression is always between 0 and SD_Y. It is zero when r is equal to plus or minus1 and SD_Y when r is equal to zero. (We can try this by substituting r is equal to 1 and r is equal to zero into the expression above.)

When r is equal to plus or minus 1, the regression line accounts for all of the variability of Y, and the root mean square of the vertical residuals is zero.

When r is equal to zero, the regression line does not "explain" any of the variability of Y: The regression line is a horizontal line at height mean(Y), so the root mean square of the vertical residuals from the regression line is the root mean square of the deviations of the values of Y from the mean of Y, which is, by definition, the SD of Y.

When r is not zero, the regression line accounts for some of the variability of Y, so the scatter around the regression line is less than the overall scatter in Y.

If the scatter plot is football-shaped, the mean of the values in a thin vertical strip will be about the same as the height of the regression line, and the SD of the values in a vertical strip will be about the same as the root mean square (that is vertical) error of regression.

The regression line is a smoothed version of the graph of averages. The height of the regression line at the point *x* is an estimate of the average of the values of Y for individuals whose value of X is close to *x*. If the scatter plot is football – shaped, the regression line follows the graph of averages reasonably well.

In each vertical slice, the deviations of the values of Y from their mean are approximately the vertical residuals of those values of Y from the regression line. The SD of the values of Y in the slice is thus approximately the root mean square of the residuals in the slice. Because football-shaped scatter plots are homescedastic, the SD of the values of Y in every vertical slice is about the same, so the root mean square error of regression is a reasonable estimate of the scatter of the values of Y in vertical slices through football-shaped scatter plots.

In contrast, when the scatter plot is not football-shaped—because of nonlinearity, heteroscedasticity or outliers—the root mean square error of regression is not a good measure of the scatter in a "typical" vertical slice.

If a scatter plot is homoscedastic and shows nonlinear association, the root mean square error of regression tends to overestimate the scatter in a typical vertical slice: the residuals have a contribution from scatter around the average in the slice, and a contribution from the difference

between the average in the slice and the height of the regression line in the slice.

Similarly, if a scatter plot is heteroscedastic and shows linear association, the root mean square error of regression will overestimate the scatter in some slices and underestimate the scatter in other slices.

If a scatter plot has outliers, is otherwise homoscedastic, and shows linear association, the root mean square error of regression will tend to overestimate the scatter in slices. The strength of linear association affects the size of the root mean square error of regression, but it does not affect whether the root mean square error of regression is a good estimate of the scatter in vertical slices.

Here's a summary of our learning in this session:

- The meaning of regression
- Application of regression
- Comparison of regression with correlation
- Importance or utility of regression analysis
- Methods of studying regression
- Error in Regression
- Root mean square error