Government Arts College for Women, Salem-8

Department of Economics

Class: III.B.A.Economics Subject:Basic Econometrics

INTRODUCTION

UNIT-I

Data can be defined as a collection of facts or information from which conclusions may bedrawn. Data may be qualitative or quantitative. Once we know the difference between them, we can know how to use them.

Qualitative Data: They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative. They are more exploratory than conclusive in nature.

Quantitative Data: These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them. For example, data on the number of students playing different sports from your class gives an estimate of how many of the total students play which sport. This information is numerical and can be classified as quantitative.

Discrete Data: These are data that can take only certain specific values rather than a range of values. For example, data on the blood group of a certain population or on their genders is termed as discrete data. A usual way to represent this is using bar charts.

Continuous Data: These are data that can take values between a certain range with the highest and lowest values. The difference between the highest and lowest value is called the range of data. For example, the age of persons can take values even in decimals or so is the case of the height and weights of the students of your school. These are classified as continuous data. Continuous data can be tabulated in what is called a frequency distribution. They can be graphically represented using histograms. Depending on the source, it can classify as primary data or secondary data. Let us take

a look at them both.

Primary Data: These are the data that are collected for the first time by an investigator for a Specific purpose. Primary data are 'pure' in the sense that no statistical operations have been performed on them and they are original. An example of primary data is the Census of India.

Secondary Data: They are the data that are sourced from someplace that has originally collected it. This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. This information is impure as statistical operations may have been performed on them already. An example is information available on the Government of India, Department of Finance's website or in other repositories, books, journals, etc.

Collection of Primary Data

Primary data is collected in the course of doing experimental or descriptive research by doing
experiments, performing surveys or by observation or direct communication with respondents.
Several methods for collecting primary data are given below:

1. Observation Method

It is commonly used in studies relating to behavioural science. Under this method observation
becomes a scientific tool and the method of data collection for the researcher, when it serves a formulated research purpose and is systematically planned and subjected to checks and controls.

(a) **Structured (descriptive) and Unstructured (exploratory) observation**: When a observation is characterized by careful definition of units to be observed, style of observer, conditions for observation and selection of pertinent data of observation it is a structured observation. When there characteristics are not

thought of in advance or not present it is a unstructured observation.

(b) **Participant, Non-participant and Disguised observation:** When the observer observes by making himself more or less, the member of the group he is observing, it is participant observation but when the observer observes by detaching him from the group under observation it is non participant observation. If the observer observes in such a manner that his presence is unknown to the people he is observing it is disguised observation.

(c) **Controlled (laboratory) and Uncontrolled (exploratory) observation:** If the observation ,takes place in the natural setting it is a uncontrolled observation but when observer takes place according to some pre-arranged plans, involving experimental procedure it is a controlled observation.

Advantages

- Subjective bias is eliminated.
- Data is not affected by past behaviour or future intentions
- Natural behaviour of the group can be recorded

Limitations

- Expensive methodology
- Information provided is limited
- Unforeseen factors may interfere with the observational task

Interview Method

This method of collecting data involves presentation of oral verbal stimuli and reply in terms of oral - verbal responses. It can be achieved by two ways:

- **Personal Interview:** It requires a person known as interviewer to ask questions generally in a face to face contact to the other person. It can be:

- **Direct personal investigation:** The interviewer has to collect the information personally from the services concerned.

- **Indirect oral examination:** The interviewer has to cross examine other persons who are suppose to have a knowledge about the problem.

- **Structured Interviews:** Interviews involving the use of pre- determined questions and of highly standard techniques of recording.
- **Unstructured interviews:** It does not follow a system of pre-determined questions and is characterized by flexibility of approach to questioning.
- **Focused interview:** It is meant to focus attention on the given experience of the respondent and its effect. The interviewer may ask questions in any manner or sequence with the aim to explore reasons and motives of the respondent.
- **Clinical interviews:** It is concerned with broad underlying feeling and motives or individual's life experience which are used as method to elicit information under this method at the interviewer direction.
- **Non directive interview:** The interviewer's function is to encourage the respondent to talk about the given topic with a bare minimum of direct questioning.

Advantages:

- More information and in depth can be obtained
- Samples can be controlled
- There is greater flexibility under this method
- Personal information can as well be obtained
- Mis-interpretation can be avoided by unstructured interview.

Limitations

- It is an expensive method
- Possibility of bias interviewer or respondent
- More time consuming
- Possibility of imaginary info and less frank responses
- High skilled interviewer is required

(B) Telephonic Interviews: It requires the interviewer to collect information by contacting
respondents on telephone and asking questions or opinions orally.

Advantages:

- It is flexible, fast and cheaper than other methods
- Recall is easy and there is a higher rate of response
- No field staff is required.

Limitations:

- Interview period exceed five minutes maximum which is less
- Restricted to people with telephone facilities
- Questions have to be short and to the point
- Less information can be collected.

Questionnaire

In this method a questionnaire is sent (mailed) to the concerned respondents who are expected to read, understand and reply on their own and return the questionnaire. It consists of a number of questions printed on typed in a definite order on a form on set of forms. It is advisable to conduct a `Pilot study' which is the rehearsal of the main survey by experts for testing the questionnaire for weaknesses of the questions and techniques used.

Essentials of a good questionnaire:

- It should be short and simple
- Questions should proceed in a logical sequence
- Technical terms and vague expressions must be avoided.

- Control questions to check the reliability of the respondent must be present
- Adequate space for answers must be provided
- Brief directions with regard to filling up of questionnaire must be provided
- The physical appearances – quality of paper, colour etc must be good to attract the
- attention of the respondent

Advantages:

- Free from bias of interviewer
- Respondents have adequate time to give
- Respondents have adequate time to give answers
- Respondents are easily and conveniently approachable
- Large samples can be used to be more reliable

Limitations:

- Low rate of return of duly filled questionnaire
- Control over questions is lost once it is sent
- It is inflexible once sent
- Possibility of ambiguous or omission of replies
- Time taking and slow process

4. Schedules

This method of data collection is similar to questionnaire method with the difference that schedules are being filled by the enumerations specially appointed for the purpose. Enumerations explain the aims and objects of the investigation and may remove anyn misunderstanding and help the respondents to record answer. Enumerations should be well trained to perform their job; he/she should be honest hard working and patient. This type of data is helpful in extensive enquiries however it is very expensive.

Collection of Secondary Data

A researcher can obtain secondary data from various sources. Secondary data may either be

published data or unpublished data. Published data are available in:

- Publications of government
- Technical and trade journals
- Reports of various businesses, banks etc.
- Public records
- Statistical or historical documents.

Unpublished data may be found in letters, diaries, unpublished biographies or work. Before using secondary data, it must be checked for the following characteristics:

- **Reliability of data:** Who collected the data? From what source? Which methods? Time? Possibility of bias? Accuracy?
- **Suitability of data:** The object, scope and nature of the original enquiry must be studies and then carefully scrutinize the data for suitability.
- **Adequacy:** The data is considered inadequate if the level of accuracy achieved in data is found inadequate or if they are related to an area which may be either narrower or wider than the area of the present enquiry.

Census and Sample of Data

In Statistics, the basis of all statistical calculation or interpretation lies in the collection of data. There are numerous methods of data collection. In this lesson, we shall focus on two primary methods and understand the difference between them. Both are suitable in different cases and the knowledge of these methods is important to understand when to apply which method. These two methods are Census method and Sampling method.

Census Method:

Census method is that method of statistical enumeration where all members of the population are studied. A population refers to the set of all observations under

concern. For example, if you want to carry out a survey to find out student's feedback about the facilities of your school, all the students of your school would form a part of the 'population' for your study. At a more realistic level, a country wants to maintain information and records about all households. It can collect this information by surveying all households in the country using the census method. In our country, the Government conducts the Census of India every ten years. The Census appropriates information from households regarding their incomes, the earning members, the total number of children, members of the family, etc. This method must take into account all the units. It cannot leave out anyone in collecting data. Once collected, the Census of India reveals demographic information such as birth rates, death rates, total population, population growth rate of our country, etc. The last census was conducted in the year 2011.

Sampling Method:

Like we have studied, the population contains units with some similar characteristics on the basis of which they are grouped together for the study. In case of the Census of India, for example, the common characteristic was that all units are Indian nationals. But it is not always practical to collect information from all the units of the population. It is a time-consuming and costly method. Thus, an easy way out would be to collect information from some representative group from the population and then make observations accordingly. This representative group which contains some units from the whole population is called the sample.

Sample Selection:

The first most important step in selecting a sample is to determine the population. Once the

population is identified, a sample must be selected. A good sample is one which is:

- Small in size.
- Provides adequate information about the whole population.

- Takes less time to collect and is less costly.

In the case of our previous example, you could choose students from your class to be the representative sample out of the population (all students in the school). However, there must be some rationale behind choosing the sample. If you think your class comprises a set of students who will give unbiased opinions/feedback or if you think your class contains students from different backgrounds and their responses would be relevant to your student, you must choose them as your sample. Otherwise, it is ideal to choose another sample which might be more relevant.

Again, realistically, the government wants estimates on the average income of the Indian household. It is difficult and time-consuming to study all households. The government can simply choose, say, 50 households from each state of the country and calculate the average of that to arrive at an estimate. This estimate is not necessarily the actual figure that would be arrived at if all units of the population underwent study. But, it approximately gives an idea of what the figure might look like.

Sampling Techniques

Sampling helps a lot in research. It is one of the most important factors which determine the accuracy of your research/survey result. If anything goes wrong with your sample then it will be directly reflected in the final result. There are lot of techniques which help us to gather sample depending upon the need and situation. This blog post tries to explain some of those techniques.

To start with, let's have a look on some basic terminology.

- **Population** is the collection of the elements which has some or the other characteristic in common. Number of elements in the population is the size of the population.
- **Sample** is the subset of the population. The process of selecting a sample is known as sampling. Number of elements in the sample is the sample size.

Sampling

There are lot of sampling techniques which are grouped into two categories as:

- Probability Sampling
- Non- Probability Sampling

The difference lies between the above two is weather the sample selection is based on randomization or not. With randomization, every element gets equal chance to be picked up and to be part of sample for study.

Probability Sampling

This Sampling technique uses randomization to make sure that every element of the population gets an equal chance to be part of the selected sample. It's alternatively known as random sampling.

Simple Random Sampling:

Every element has an equal chance of getting selected to be the part sample. It is used when we don't have any kind of prior information about the target population.

For example: Random selection of 20 students from class of 50 students. Each student has equal chance of getting selected. Here probability of selection is 1/50

Stratified Sampling

This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous and heterogeneous among the other subgroups formed. And then the elements are randomly selected from each of these strata. We need to have prior information about the population to create subgroups.

Cluster Sampling

Our entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling. Clusters are identified using details such as age, sex, location etc. Cluster sampling can be done in following ways:

Single Stage Cluster Sampling

Entire cluster is selected randomly for sampling. **Two Stage Cluster Sampling** Here first we randomly select clusters and then from those selected clusters we randomly select elements for sampling

Systematic Clustering

Here the selection of elements is systematic and not random except the first element. Elements of a sample are chosen at regular intervals of population. All the elements are put together in a sequence first where each element has the equal chance of being selected. For a sample of size n, we divide our population of size N into subgroups of k elements. We select our first element randomly from the first subgroup of k elements. To select other elements of sample, perform following:

We know number of elements in each group is k i.e N/n

So if our first element is n1 then

Second element is n1+k i.e n2

Third element n2+k i.e n3 and so on..

Taking an example of N=20, n=5

No of elements in each of the subgroups is N/n i.e 20/5 =4= k

Now, randomly select first element from the first subgroup.

If we select n1= 3

n2 = n1+k = 3+4 = 7

n3 = n2+k = 7+4 = 11

Multi-Stage Sampling

It is the combination of one or more methods described above. Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity. One or more clusters can be randomly selected from each stratum. This process continues until the cluster can't be divided anymore. For example country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.

Non-Probability Sampling

It does not rely on randomization. This technique is more reliant on the researcher's ability to select elements for a sample. Outcome of sampling might be biased and makes difficult for all the elements of population to be part of the sample equally. This type of sampling is also known as non-random sampling.

Convenience Sampling: Here the samples are selected based on the availability. This method is used when the availability of sample is rare and also costly. So based on the convenience samples are selected.

For example: Researchers prefer this during the initial stages of survey research, as it's quick and easy to deliver results.

Purposive Sampling: This is based on the intention or the purpose of study. Only those elements will be selected from the population which suits the best for the purpose of our study.

For Example: If we want to understand the thought process of the people who are interested in pursuing master's degree then the selection criteria would be "Are you interested for Masters in..?" All the people who respond with a "No" will be excluded from our sample.

Quota Sampling: This type of sampling depends of some pre-set standard. It selects the representative sample from the population. Proportion of characteristics/ trait in sample should be same as population. Elements are selected until exact proportions of certain types of data is obtained or sufficient data in different categories is collected.

For example: If our population has 45% females and 55% males then our sample should reflect the same percentage of males and females.

Referral /Snowball Sampling: This technique is used in the situations where the

population is completely unknown and rare. Therefore we will take the help from the first element which we select for the population and ask him to recommend other elements who will fit the description of the sample needed. So this referral technique goes on, increasing the size of population like a
snowball.

For example: It's used in situations of highly sensitive topics like HIV Aids where people will not openly discuss and participate in surveys to share information about HIV Aids. Not all the victims will respond to the questions asked so researchers can contact people they know or volunteers to get in touch with the victims and collect information Helps in situations where we do not have the access to sufficient people with the characteristics we are seeking. It starts with finding people to study.

UNIT-II

Concept of t, Chi Square and F Distribution

The t distribution

The probability distribution that will be used most of the time in this book is the so called f distribution. The f-distribution is very similar in shape to the normal distribution but works better for small samples. In large samples the f-distribution converges to the normal distribution.

Properties of the t-distribution:

In the previous section we explained how we could transform a normal random variable with an arbitrary mean and an arbitrary variance into a standard normal variable. That was under condition that we knew the values of the population parameters. Often it is not possible to know the population variance, and we have to rely on the sample value. The transformation formula would then have a distribution that is different from the normal in small samples. It would
instead be f-distributed. Assume that you have a sample of 60 observations and you found that the sample mean equals 5 and the sample variance equals 9. You would like to know if the population mean is different from 6. We state the following hypothesis:
H0: u= 6 H1 : / * 6

We use the transformation formula to form the test function

1. The f-distribution is symmetric around its mean.
2. The mean equals zero just as for the standard normal distribution.
3. The variance equals k/(k-2), with k being the degrees of freedom.

Example:

Observe that the expression for the standard deviation contains an S. S represents the sample standard deviation. Since it is based on a sample it is a random variable, just as the mean. The test function therefore contains two random variables. That implies more variation, and therefore a distribution that deviates from the standard normal. It is possible to show that the distribution of this test function follows the ^-distribution with n-1 degrees of freedom, where n is the sample size. Hence in our case the test value equals: The test value has to be compared with a critical value. If we choose a significance level of 5% the critical values according to the ^-distribution would be [-2.0; 2.0]. Since the test value is located outside the interval we can say that we reject the null hypothesis in favor for the alternative hypothesis. That we have no information about the population mean is of no problem, because we assume that the population mean takes a value according to the null hypothesis. Hence, we assume that we know the true population mean. That is part of the test procedure.

The Chi-square distribution

Until now we have talked about the population mean and performed tests related to the mean. Often it is interesting to make inference about the population variance as well. For that purpose we are going to work with another distribution, the Chi-square distribution. Statistical theory shows that the square root of a standard normal variable is distributed according to the Chisquare distribution and it is denoted x2, and has one degree of freedom. It turns out that the sum of squared independent standard normal variables also is Chi-squared distributed. We have:

Properties of the Chi-squared distribution:

1. The Chi-square distribution takes only positive values
2. It is skewed to the right in small samples, and converges to the normal distribution as the degrees of freedom goes to infinity

3. The mean value equals k and the variance equals 2k, where k is the degrees of freedom

In order to perform a test related to the variance of a population using the sample variance we need a test function with a known distribution that incorporates those components. In this case we may rely on statistical theory that shows that the following function would work: Where S2 represents the sample variance, Sigma Square the population variance, and n-1 the degrees of freedom used to calculate the sample variance. How could this function be used to perform a test related to the population variance?

Example:

We have a sample taken from a population where the population variance a given year was — =

400. Some years later we suspect that the population variance has increased and would like test if

that is the case. We collect a sample of 25 observations and state the following hypothesis:

Using the 25 observations we found a sample variance equal to 600. Using this information we

set up the test function and calculate the test value:

We choose a significance level of 5% and find a critical value in Table A3 equal to 36.415. Since

the test value is lower than the critical value we cannot reject the null hypothesis. Hence we

cannot say that the population variance has changed.

The F-distribution

The final distribution to be discussed in this chapter is the F-distribution. In shape it is very similar to the Chi-square distribution, but is a construction based on a ratio of two independent Chi-squared distributed random variables. An F-distributed random variable therefore has two sets of degrees of freedom, since each variable in this ratio has its own degrees of freedom. That is:

Properties of the F-distribution:

1. The F-distribution is skewed to the right and takes only positive values

2. The F-distribution converges to the normal distribution when the degrees of freedom become large

3. The square of a f-distributed random variable with k degrees of freedom become Fdistributed:

tk = F] £ The P-distribution can be used to test population variances. It is especially interesting when we would like to know if the variances from two different populations differ from each other. Statistical theory says that the ratio of two sample variances forms an P-distributed random variable with n1 -1 and n2 -1 degrees of freedom:

Example:

Assume that we have two independent populations and we would like to know if their variances are different from each other. We therefore take two samples, one from each population, and form the following hypothesis: Using the two samples we calculate the sample variances, Sj2 = 8.38 and S% = 13.14 with nl = 26 and "2 = 30. Under the null hypothesis we know that the ratio of the two sample variances is P-distributed with 25 and 29 degrees of freedom. Hence we form the test function and calculate the test value:

This test value has to be compared with a critical value. Assume that we choose a significance level of 5%. Using Table A4 in the appendix, we have to find a critical value for a two sided test. Since the area outside the interval should sum up to 5%, we must find the upper critical point that corresponds to 2.5%. If we look for that value in the table we find 2.154. We call this upper point F0.025. In order to find the lover point we can use the following formula: We have therefore received the following interval: [0.464; 2.154]. The test value lies within this interval, which means that we are unable to reject the null hypothesis. It is therefore quite possible that the two population variances are the same.

WHAT IS ECONOMETRICS?

Econometrics is the quantitative application of statistical and mathematical models using data to develop theories or test existing hypotheses in economics, and for

forecasting future trends from historical data. It subjects real-world data to statistical trials and then compares and contrasts the results against the theory or theories being tested. Depending on if you are interested in testing an existing theory or using existing data to develop a new hypothesis based on those observations, econometrics can be subdivided into two major categories: theoretical and applied. Those who routinely engage in this practice are commonly known as econometricians. Econometrics deals with the measurement of economic relationships. It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of mathematical form of the economic relationships. The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods. The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations. It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

Econometric Model

A model is a simplified representation of a real world process. It should be representative in the sense that it should contain the salient features of the phenomena under study. In general, one of the objectives in modeling is to have a simple model to explain a complex phenomenon. Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic. In practice, generally all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model. Rest of the variables are dumped in a basket called “disturbances” where the disturbances are random variables. This is the main

difference between the economic modeling and econometric modeling. This is also the main difference between the mathematical modeling and statistical modeling. The mathematical modeling is exact in nature whereas the statistical modeling contains a stochastic term also. An economic model is a set of assumptions that describes the behaviour of an economy, or more general, a phenomenon. An econometric model consists of:

- A set of equations describing the behaviour. These equations are derived from the economic model and have two parts – observed variables and disturbances.
- A statement about the errors in the observed values of variables.
- A specification of the probability distribution of disturbances.

Aims of Econometrics

1. **Formulation and specification of econometric models:** The economic models are formulated in an empirically testable form. Several econometric models can be derived from an economic model. Such models differ due to different choice of functional form, specification of stochastic structure of the variables etc.
2. **Estimation and testing of models:** The models are estimated on the basis of observed set of data and are tested for their suitability. This is the part of statistical inference of the modeling. Various estimation procedures are used to know the numerical values of the unknown parameters of the model. Based on various formulations of statistical models, a suitable and appropriate model is selected.
3. **Use of models:** The obtained models are used for forecasting and policy formulation which is an essential part in any policy decision. Such forecasts help the policy makers to judge the goodness of fitted model and take necessary measures in order to re-adjust the relevant economic variables.

Econometrics and Statistics

Econometrics differs both from mathematical statistics and economic statistics. In economic statistics, the empirical data is collected recorded, tabulated and used in

describing the pattern in their development over time. The economic statistics is a descriptive aspect of economics. It does not provide either the explanations of the development of various variables or measurement of the parameters of the relationships. Statistical methods describe the methods of measurement which are developed on the basis of controlled experiments. Such methods may not be suitable for economic phenomenon as they don't fit in the framework of controlled experiments. For example, in real world experiments, the variables usually change continuously and simultaneously and so the set up of controlled experiments are not suitable. Econometrics uses statistical methods after adapting them to the problems of economic life.

These adopted statistical methods are usually termed as econometric methods. Such methods are adjusted so that they become appropriate for the measurement of stochastic relationships. These adjustments basically attempts to specify attempts to the stochastic element which operate in real world data and enters into the determination of observed data. This enables the data to be called as random sample which is needed for the application of statistical tools. The theoretical econometrics includes the development of appropriate methods for the measurement of economic relationships which are not meant for controlled experiments conducted inside the laboratories. The econometric methods are generally developed for the analysis of non-experimental data. Whereas, the applied econometrics includes the application of econometric methods to specific branches of econometric theory and problems like demand, supply, production, investment, consumption etc. The applied econometrics involves the application of the tools of econometric theory for the analysis of economic phenomenon and forecasting the economic behaviour.

Econometrics and Regression analysis

One of the very important roles of econometrics is to provide the tools for modeling on the basis of given data. The regression modeling technique helps a lot in this task. The regression models can be either linear or non-linear based on which we have linear regression analysis and non-linear regression analysis. We will consider only the tools of linear regression analysis and our main interest will be the fitting of linear regression

model to a given set of data.

UNIT-III

Regression versus Correlation

Correlation:

The term correlation is a combination of two words 'Co' (together) and relation (connection) between two quantities. Correlation is when, at the time of study of two variables, it is observed that a unit change in one variable is retaliated by an equivalent change in another variable, i.e. direct or indirect. Or else the variables are said to be uncorrelated when the movement in one variable does not amount to any movement in another variable in a specific direction. It is a statistical technique that represents the strength of the connection between pairs of variables. Correlation can be positive or negative. When the two variables move in the same direction, i.e. an increase in one variable will result in the corresponding increase in another variable and vice versa, then the variables are considered to be positively correlated. For instance: profit and investment. On the contrary, when the two variables move in different directions, in such a way that an increase in one variable will result in a decrease in another variable and vice versa then this situation is known as negative correlation. For instance: Price and demand of a product.

Regression: A statistical technique for estimating the change in the metric dependent variable due to the change in one or more independent variables, based on the average mathematical relationship between two or more variables is known as regression. It plays a significant role in many human activities, as it is a powerful and flexible tool which used to forecast the past, present or future events on the basis of past or present events. For instance: On the basis of past records, a business's future profit can be estimated. In a simple linear regression, there are two variables x and y, wherein y depends on x or say influenced by x. Here y is called as dependent, or criterion variable and x is independent or predictor variable. The regression line of y on x is expressed as under:

$y = a + bx$ Where, a = constant and b = regression coefficient. In this equation, a and b are two regression parameters.

Differences between Correlation and Regression

The points given below, explains the difference between correlation and regression in detail:

- A statistical measure which determines the co-relationship or association of two quantities is known as Correlation. Regression describes how an independent variable is numerically related to the dependent variable.
- Correlation is used to represent the linear relationship between two variables. On the contrary, regression is used to fit the best line and estimate one variable on the basis of another variable.
- In correlation, there is no difference between dependent and independent variables i.e. correlation between x and y is similar to y and x. Conversely, the regression of y on x is different from x on y.
- Correlation indicates the strength of association between variables. As opposed to, regression reflects the impact of the unit change in the independent variable on the dependent variable.
- Correlation aims at finding a numerical value that expresses the relationship between variables. Unlike regression whose goal is to predict values of the random variable on the basis of the values of fixed variable.

Correlation and Regression are the two analysis based on multivariate distribution. A multivariate distribution is described as a distribution of multiple variables. Correlation is described as the analysis which lets us know the association or the absence of the relationship between two variables ‘x’ and ‘y’. On the other end, Regression analysis, predicts the value of the dependent variable based on the known value of the independent variable, assuming that average mathematical relationship between two or more variables.

The difference between correlation and regression is one of the commonly asked questions in interviews. Moreover, many people suffer ambiguity in understanding these two. So, take a full read of this article to have a clear understanding on these two.

Regression Model

The classical linear regression model can be expressed as follows equation, where Yi is dependent variable, Xi is the independent or explanatory variable, α is the

regression constant or intercept, β is the regression coefficient for the effect of Xi on Yi or slope of the regression equation, and ei is the error we make in predicting Yi from Xi.
$Yi = \alpha + \beta Xi + ei$

Steps in Regression Analysis

- Statement of the problem under consideration
- Choice of relevant variables
- Collection of data on relevant variables
- Specification of model
- Choice of method for fitting the data
- Fitting of model
- Model validation and criticism

Types of Data for Regression as well as Econometrics Analysis

Cross section data: Cross section data give information on the variables concerning individual agents (e.g., consumers or produces) at a given point of time. For example, data on income across sample individuals for a particular point of time say in the year 2015.

Time series data: Time series data give information about the numerical values of variables from period to period and are collected over time. For example, the data during the years 1990-
2010 for monthly income constitutes a time series data.

Panel data: The panel data are the data from repeated survey of a single (cross-section) sample in different periods of time.

AUTOCORRELATION

One of the assumptions of the classical linear regression model is that the disturbance or errorterm of the model is independent. Symbolically, it means that, for

the model:

$$Yt = \alpha + \beta Xt + et$$

$$\text{Covariance } (et, es) = 0 \text{ for } t \neq s$$

This feature of regression disturbance is known as serial independence or non-autocorrelation, which replies that the value of disturbance term in one period is not correlated with its value in another period. Violation of this assumption, arises mainly in case of time series data, is called as autocorrelation. So, autocorrelation is just as correlation measures the extent of a linear relationship between two variables and it measures the linear relationship between lagged values of a time series. It is a characteristic of data which shows the degree of similarity between the values of the same variables over successive time intervals. This post explains what autocorrelation is, types of autocorrelation - positive and negative autocorrelation, as well as how to diagnose and test for auto correlation. When you have a series of numbers, and there is a pattern such that values in the series can be predicted based on preceding values in the series, the series of numbers is said to exhibit autocorrelation. This is also known as serial correlation and serial dependence. The existence of autocorrelation in the residuals of a model is a sign that the model may be unsound. Autocorrelation is diagnosed using a correlogram (ACF plot). There is a very popular test called the Durbin Watson test that detects the presence of autocorrelation. If the researcher detects autocorrelation in the data, then the first thing the researcher should do is to try to find whether or not it is pure. If it is pure, then one can transform it into the original model that is free from pure autocorrelation. In presence of the autocorrelation in data, the ordinary least square (OLS) estimation technique can’t be applied as the estimate violate the BLUE property. The auto part of autocorrelation is from the Greek word for self, and autocorrelation means data that is correlated with itself, as opposed to being correlated with some other data. Consider the nine values of Y below. The column to the right shows the last eight of these values, moved “up” one row, with the first value deleted. When we correlate these two columns of data, excluding the last observation that has missing values, the correlation is 0.64. This means that the data is correlated with itself (i.e., we have autocorrelation/serial correlation).