Basic Econometrics
ECOPG-203: ECONOMETRICS
Prepared by: Sidhartha Sankar Laha

(The lecture note is constructed on the basis of collections from several notes, books, journals and websites. In case of any difficulty to understand one may contact the instructor. Useful disclaimers apply)

SYLLABUS

Unit-I: Classical Linear Regression Model

Collection of Data: Primary Data, Secondary Data, Various Methods of Collection of Primary Data, Census and Sample of Data, Various Sampling Techniques, t Distribution , Chi Square and F Distribution, What is a Regression Model, Regression versus Correlation, Simple Regression, The Assumption underlying the Classical Linear Regression Model, Properties of OLS estimator, introduction to Statistical inference.

Unit-II: Further Development and Analysis of the Classical Linear Regression Model

Generalising the Simple model to Multiple Linear Regression Model, The Constant term, how are the Parameters calculated in the generalized case? Testing multiple hypothesis: the F-test, Goodness of fit statistics

Unit-III: Autocorrelation

Nature and Scope of Autocorrelation, Sources of Autocorrelation- Omitted Explanatory variables, Interpolation in the statistical observations, Teasting for Autocorrelation, Soluations for Autocorrelation, Methods of estimating ρ - Estimated from the Residuals, D-W Statistics, Theil and Nagar Method.

Unit-IV: Multicollinearity

Consequences of Multicollinearity, Direction of Multicollinearity, Auxiliary Regressions, Variance inflation factor, Relationship between R^2 and VIF, Solution to the problem of Multicollinearity.

Unit-V: Heteroscedasticity

Consequences of Heteroscedasticity, Testing of Heteroscedasticity with error term, Causes of Heteroscedasticity, Test for Heteroscedasticity- Spearmans rank correlation test, Park test, Glejser test, Goldfeld- Quandt test, Breusch- Pagan Test, Remedies for Heteroscedasticity.

Unit-VI: Identification and Simultaneous Equation Methods

Endogenous and exogenous variables, structural and reduced form equations, structural and reduced- from equations, Identification problems in economics, Order and rank for identifiability, Simultaneous versus recursive equation system, 2SLS- exactly and over indentified, Recursive methods and OLS; Indirect least squares (ILS); 2SLS, 3SLS and ML methods – applications.

1. INTRODUCTION

Data can be defined as a collection of facts or information from which conclusions may be drawn. 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

2. 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:

(A) 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.

3. 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 any 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.

Difference between Census and Sample Surveys

Parameter	Census	Sample Survey
Definition	A statistical method that studies all the units or members of a	A statistical method that studies only a representative group of the population, and

	population.	not all its members.
Calculation	Total/Complete	Partial
Time involved	It is a time-consuming process.	It is a quicker process.
Cost involved	It is a costly method.	It is a relatively inexpensive method.
Accuracy	The results obtained are accurate as each member is surveyed. So, there is a negligible error.	The results are relatively inaccurate due to leaving out of items from the sample. The resulting error is large.
Reliability	Highly reliable	Low reliability
Error	Not present	The smaller the sample size, the larger the error.
Relevance	This method is suited for heterogeneous data.	This method is suited for homogeneous data.

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.

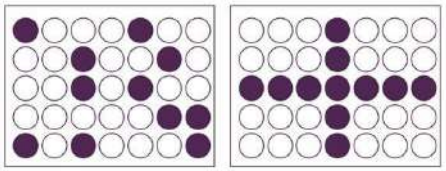
Population is the collection of the elements which has some or the other characteristic in common. Number of elements in the population is the size of the population.

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

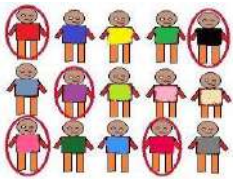

Probability Sampling Vs Non-Probability Sampling

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.

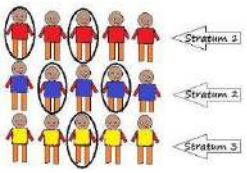
Simple Random Sampling: Every element has an equal chance of getting selected to be the part sample. It is used when we don't have any kind of prior information about the target population.

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



Single Random Sampling

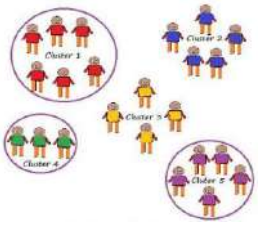
Stratified Sampling
This technique divides the elements of the population into small subgroups (strata) based on the similarity in such a way that the elements within the group are homogeneous and heterogeneous among the other subgroups formed. And then the elements are randomly selected from each of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



Stratified Sampling

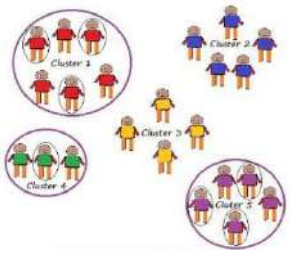
Cluster Sampling
Our entire population is divided into clusters or sections and then the clusters are randomly selected. All the elements of the cluster are used for sampling. Clusters are identified using details such as age, sex, location etc. Cluster sampling can be done in following ways:

Single Stage Cluster Sampling
Entire cluster is selected randomly for sampling.



Single Stage Cluster Sampling

Two Stage Cluster Sampling
Here first we randomly select clusters and then from those selected clusters we randomly select elements for sampling



Two Stage Cluster Sampling

Systematic Clustering

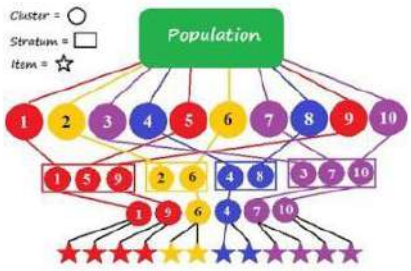
Here the selection of elements is systematic and not random except the first element. Elements of a sample are chosen at regular intervals of population. All the elements are put together in a sequence first where each element has the equal chance of being selected.

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

Systematic Clustering

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.



Multi-Stage Sampling

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.

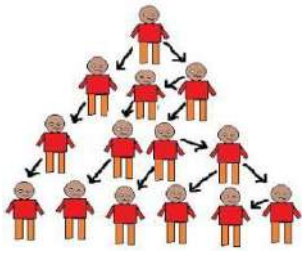
Purposive Sampling: This is based on the intention or the purpose of study. Only those elements will be selected from the population which suits the best for the purpose of our study.

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.



Referral /Snowball Sampling

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.

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:

H_0: 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:

$$t = \frac{\overline{X} - \mu_X}{S/\sqrt{n}} \sim t_{(n-1)}$$

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:

$$t = \frac{\overline{X} - \mu_X}{S/\sqrt{n}} = \frac{5-6}{3/\sqrt{60}} = -2.58$$

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 Chi-square 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:

$$Z_1^2 + Z_2^2 + ... + Z_k^2 \sim \chi^2_{(k)}$$

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:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Where S^2 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:

$$H_0 : \sigma^2 = 400$$

$$H_1 : \sigma^2 > 400$$

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:

$$\text{Test Function} = \frac{(n-1)S^2}{s^2} = \frac{(24-1)\times 600}{400} = 36$$

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:

$$\frac{\chi^2_m}{\chi^2_l} \sim F_{m,l}$$

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 F-distributed: 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 n_1 -1 and n_2 -1 degrees of freedom:

$$\frac{S_1^2}{S_2^2} \sim F_{(n_1-1)(n_2-1)}$$

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:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

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:

$$\frac{S_1^2}{S_2^2} = \frac{8.38}{13.14} = 0.638$$

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:

$$F_{0.975} = \frac{1}{F_{0.025}} = \frac{1}{2.154} = 0.464$$

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 modelling 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 modelling and econometric modelling. This is also the main difference between the mathematical modelling and statistical modelling. The mathematical modelling is exact in nature whereas the statistical modelling 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 modelling. 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 modelling on the basis of given data. The regression modelling 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.

For details one may follow Scope and Methodology of Econometrics as follows.

[Important: Chapter Page No. 3 to 10, From Point 1.2 to 1.5 (Before 1.6 at Page 10)]

1 Scope and Methodology of Econometrics

> As far as the laws of mathematics refer to reality they are not certain, and as far as they are certain they do not refer to reality.
>
> **—Albert Einstein**
> (Famous theoretical physicist who won the 1921 Nobel Prize in Physics)

> We need a special field called econometrics, and textbooks about it, because it is generally accepted that economic data possess certain properties that are not considered in standard statistics texts or are not sufficiently emphasized there for use by economists.
>
> **—Clive W.J. Granger**
> (Co-recipient of 2003 Nobel Prize in Economic Sciences)

1.1. WHAT IS ECONOMETRICS?

Econometrics is a neologism formed by combining two Greek words: *oikonomia* (meaning economics) and *metron* (meaning measure) (Tintner 1953, 33). Thus, the literal meaning of econometrics appears to be 'measurement in economics'. However, although measurement is indeed an important component of econometrics, the scope of econometrics is much wider than that. This becomes clear from various definitions of econometrics[1] provided by leading econometricians over the years. A few of those definitions are presented below.

> The method of econometric research aims, essentially, at a conjunction of economic theory and actual measurements, using the theory and technique of statistical inference as a bridge pier. (Haavelmo 1944, iii)

> [I]t is better to restrict econometrics to investigations which utilize mathematics, economics and statistics. These will be different from investigations in quantitative economics, which frequently use no mathematics. They will also be distinguished from work in mathematical economics, which is quantitative, but not empirical and uses no statistics. Finally, it will be distinct from theoretical work in statistics, which uses mathematics but is in general unrelated to economic theory. (Tintner 1953, 37)

[1] The term 'econometrics' appears to have been first used by Pawel Ciompa as early as 1910 although Ragnar Frisch is credited for coining the term and establishing it as a subject in the sense in which it is known today.

> [E]conometrics is the science which deals with the determination by statistical methods of concrete quantitative laws occurring in economic life. (Lange 1962, 13)

> Econometrics attains even broader meaning if mathematics and statistics are defined in the following broad sense: mathematics teaches how to derive propositions from other propositions; statistics teaches how to derive propositions from observed facts. Mathematics would then coincide with deductive logic, and statistics with inductive logic.... Econometrics would then be simply the application of rules of logic to economics. (Marschak 1948, 1)

> [E]conometrics may be defined as the quantitative analysis of actual economic phenomenon based on the concurrent development of theory and observation, related by appropriate methods of inference. (Samuelson, Koopmans, and Stone 1954, 142)

> Econometrics may be defined as the social science in which tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena. (Goldberger 1964, 1)

> [T]he discipline in which one studies theoretical and practical aspects of applying statistical methods to economic data for the purpose of testing economic theories (represented by carefully constructed models) and of forecasting and controlling the future path of economic variables. (Sowey 1983, 257)

> Econometrics is the field of economics that concerns itself with the application of mathematical statistics and the tools of statistical inference to the empirical measurement of relationships postulated by economic theory. (Greene 2003, 1)

> Broadly speaking, econometrics aims to give empirical content to economic relations for testing economic theories, forecasting, decision making, and for ex post decisions/policy evaluation. (Geweke, Horowitz, and Pesaran 2008, 609)

It clearly emerges from above definitions that econometrics as a branch of economics differs from other branches like mathematical economics, economic statistics, and mathematical statistics. While mathematical economics expresses economic theory in formal algebraic language without bothering about measurability or empirical verification of the theory, the econometrician often uses mathematical equations proposed by the mathematical economist but puts these equations in a form so that they lend themselves to empirical testing. The economic statistics is concerned with collecting, processing, and presenting economic data in the form of charts and tables; the econometrician uses these data to test economic theories. Further, mathematical statistics provides many tools used by econometrics. However, the econometrician needs different methods because of the special nature of economic data, which is that data are rarely generated through controlled experiments.

In brief, we can say that the objective of econometrics is to provide empirical content to economic theory. The three ingredients of econometrics are: economic theory, economic data, and statistical methods. Neither 'theory without measurement' nor 'measurement without theory' is sufficient to explain economic phenomenon. It is precisely their union

that is important, as emphasized by Ragnar Frisch in the editorial note to the inaugural issue of *Econometrica*, the journal of the Econometric Society, published in 1933 (Frisch 1933).

1.2. BRIEF HISTORY OF ECONOMETRICS

Although quantitative economic analysis is a good three centuries old, econometrics as a recognized branch of economics began to emerge only in the 1930s and 1940s with the foundation of the Econometric Society, the Cowles Commission in the United States, and the Department of Applied Economics (DAE) in Cambridge, England. In the initial years of its development, these institutions emphasized on the development of econometric methods.

The first major methodological debate in econometrics appears to have taken place over the issue of applicability of the probability calculus and the sampling theory to the analysis of economic data. Frisch himself was highly sceptical of the usefulness of these tools as his primary concern was the problems of 'multicollinearity' and 'measurement errors' which he believed were pervasive in economics. However, his approach was not accepted by the econometricians at large. Instead, it was the probabilistic rationalizations of regression analysis that formed the basis of modern econometrics. In fact, acceptance of Haavelmo's (1944) probabilistic approach marked the beginning of a new era in econometrics, and paved the way for its rapid development with the likelihood method gaining importance as a tool for identification, estimation, and inference in econometrics.

In the initial years, the researchers at the Cowles Commission were mostly busy in providing a formal solution to the problems of 'identification' and developing methods for estimation of simultaneous equations models. On the other hand, the researchers at the DAE were mostly engaged in understanding various problems of time series data and obtaining a satisfactory solution to the problem of 'spurious correlation'. Their contributions during that period helped to better analyse the economic time series data and eventually laid down the basis of what is now known as the 'time series econometrics' approach.

Several other areas where econometrics witnessed significant developments over the years are dynamic specification, latent variables, expectation formation, limited dependent variables, discrete choice models, random coefficient models, disequilibrium models, non-linear estimation, panel data models, forecasting and forecast evaluation, non-parametric and semi-parametric estimation, bootstrapping, programme evaluation methods, integration and simulation methods, Bayesian econometrics, and so on.

All in all, the development of econometrics and the impact it has made on both theoretical and empirical researches in economics over the past seven to eight decades have been phenomenal.[2] This is corroborated by the fact that already six volumes of *Handbook of Econometrics* (Engle and McFadden 1994; Griliches and Intriligator 1983, 1984, 1986; Heckman and Leamer 2001, 2007) have been published, which include as many as 77

[2] Spanos (2006) provides a retrospective, but self-critical, account of the development of econometrics.

chapters covering diverse aspects of theoretical and empirical issues in econometrics. Another dazzling fact is that out of 44 Nobel Memorial Prizes in Economic Sciences given to 71 scholars till 2012, 5 awards have gone to 8 scholars for their contributions to the field of econometrics[3] (see Table 1.1 for the list of awardees), which is second highest among all branches of economics.[4] Thus, Geweke et al. (2008, 631) rightly observed:

> Econometrics has come a long way over a relatively short period. Important advances have been made in the compilation of economic data and in the development of concepts, theories and tools for the construction and evaluation of a wide variety of econometric models. Application of econometric methods can be found in almost every field of economics.... In both theory and practice econometrics has already gone well beyond what its founders envisaged.

Table 1.1 List of Nobel Laureates from Econometrics and the Rationale behind the Award

Year	Laureate	Country	Rationale
1969	Ragnar Frisch Jan Tinbergen	Norway Netherlands	For having developed and applied dynamic models for the analysis of economic processes.
1980	Lawrence Klein	United States	For the creation of econometric models and the application to the analysis of economic fluctuations and economic policies.
1989	Trygve Haavelmo	Norway	For his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures.
2000	James J. Heckman	United States	For his development of theory and methods for analysing selective samples.
	Daniel L. McFadden	United States	For his development of theory and methods for analysing descrete choice.
2003	Robert F. Engle	United States	For methods of analysing economic time series with time-varying volatility (ARCH)*.
	Clive W.J. Granger	United Kingdom	For methods of analysing economic time series with common trends (cointegration).

* See Chapter 10, Section 10.9, for the meaning of ARCH.

1.3. METHODOLOGY OF ECONOMETRICS

By methodology of econometrics, we mean the steps which are followed in an econometric study. Broadly speaking, an econometric analysis proceeds along the following steps.

(i) The statement of economic theory or formation of hypothesis
(ii) Specification of the econometric model to test the theory or hypothesis
(iii) Estimation of parameters of the specified model

[3] Apart from these scholars, the Nobel Prize for 2011 was awarded to Thomas J. Sargent and Christopher A. Sims of the United States 'for their empirical research on cause and effect in the macro-economy'. It is to be noted that in their empirical research on macro-economics, they used econometric tools innovatively. Moreover, Christopher A. Sims advanced the 'structural VAR' which many experts consider as an extremely significant contribution in the field of macroeconometric modelling.

[4] The highest number of awards (seven) has been bagged by the macroeconomists.

(iv) Verification or statistical inference
(v) Forecasting and policy formulation

Hypothesis

Hypothesis is that aspect of economic theory which is to be tested for empirical validity. Suppose we want to examine validity of the Keynesian consumption theory that postulates that consumption is a function of income. So if we frame the statement 'consumption is a function of income', it represents a hypothesis.

Model Specification

The model is an algebraic representation of a real world process. At the stage of model specification, we decide on the precise form of functional relationship between consumption and income. For this, we may take guidance from the mathematical economist. Suppose the mathematical economist suggests the following form of relationship between consumption and income

$$Y_i = \alpha + \beta X_i \tag{1.1}$$

where

Y = consumption and X = income. The subscript i refers to the case of a particular individual ($i = 1, 2, \ldots, n$).

Two features of the above relationship become apparent.

(i) There is *one-way causation* between consumption and income. This means that consumption changes with change in income and not the other way around; and
(ii) The relationship between consumption and income is *exact* and *deterministic*. This means that given the values of parameters α and β, we obtain only one value of Y for each value of X.

However, the reality is that the relationship between consumption and income is not *exact* but a *typical* one. This becomes clear when we look at the data which show that for each value of X (income) there is a whole distribution of the values of Y (consumption). In other words, persons with same income level are found to have different levels of consumption. This type of a situation calls for a stochastic specification of the consumption function, which is done by writing our model as

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{1.2}$$

Here ε_i is called the *stochastic term* or *disturbance term*, or *error term*. Equation (1.2) represents an econometric specification of the consumption–income relationship as against mathematical specification of such relationship provided by (1.1). The consumption–income relationship represented by (1.2) is also stochastic because of inclusion of the stochastic term ε_i. The term 'stochastic' here means that for each value of X, there is a whole

distribution of the values of Y so that it is not possible to forecast the value of Y exactly. This uncertainty concerning Y arises precisely because of the presence of stochastic term ε_i. Since ε_i is a random variable, Y_i in (1.2) is also a random variable.

The above discussion suggests that we make stochastic specification of relationship between the variables in econometrics, which is possible only with the inclusion of the stochastic or disturbance term ε_i in our model. Otherwise, the relationship will continue to remain *exact* or *deterministic*. In fact, inclusion of the disturbance term or our preference for a stochastic specification may be justified at least in three important ways (Kennedy 2008, 3).

(i) *Human indeterminacy*: It is believed that human behaviour is such that actions taken even under ideal circumstances will differ in a random way. In that case, the disturbance term helps to capture random or unpredictable behaviour of human beings.

(ii) *Influence of omitted variables*: Although income might be the major determinant of consumption, it is not the only determinant. Other variables like individuals' caste, sex, education, liquid asset holdings, etc., may also have a systematic influence on their consumption levels. However, in model (1.1), we have explicitly recognised income as the only determinant of consumption although we cannot rule out the possibility of these 'omitted' factors determining individuals' consumption levels. It is here the disturbance term helps by capturing the net influence of such omitted factors/variables on the dependent variable of our model, which is consumption.

(iii) *Measurement error*: It is possible that the variable being explained (which is consumption in our example) cannot be measured accurately in some cases either because of data collection difficulties or because of it being unmeasurable. In such a situation, the disturbance term can be thought of representing the measurement errors.

Estimation

Our objective here is to obtain estimates or numerical values of the unknown parameters of model (1.2) by using any one of the estimation techniques. Some popular estimation techniques used by the econometricians are Ordinary Least Squares (OLS), Maximum Likelihood (ML), Moment, etc.

Verification or Inference

In this stage, we develop suitable criteria to examine whether the estimates obtained are in conformity with the expectations of the theory that is being tested. For this purpose, we depend on the branch of statistical theory known as statistical inference.

Forecasting and Policy Formulation

This is the final stage where we utilize the estimated model for the purpose of forecasting and/or policy formulation.

1.4. NECESSARY ASSUMPTIONS FOR ESTIMATION

Once we have specified the behavioural relationship between Y and X as in model (1.2), our next task is to estimate the values of two unknown parameters of our model which are α and β. For this purpose, we need data on variables Y and X and also the disturbance term ε. But the difficulty is that ε is not observable like Y and X. Hence, to estimate the above model, we guess the values of ε. In other words, we make some reasonable assumptions about the shape of the distribution of each ε. Specifically, we make the following assumptions.

(i) *The mean or expected value of disturbance term ε is zero.* This is algebraically expressed by writing

$$E(\varepsilon_i) = 0 \text{ for all } i$$

This assumption means that for any given X, ε may have different values, but on an average it is zero. This assumption is important in that violation of it leads to the 'biased intercept' problem.[5]

(ii) *The disturbances have uniform variance* which is known as the assumption of *homoskedasticity*. Algebraically,

$$\begin{aligned} Var(\varepsilon_i) &= E[\varepsilon_i - E(\varepsilon_i)]^2 \\ &= E(\varepsilon_i^2) \qquad [\because E(\varepsilon_i) = 0] \\ &= \sigma^2 \text{ constant for all } i \end{aligned}$$

In plain language, this assumption means that every disturbance has the same variance, which is unknown. It also implies that the variance of ε would not be higher for higher values of X than for lower values. The violation of this assumption creates an econometric problem called *heteroskedasticity*.

(iii) The *disturbances are uncorrelated* which is known as the assumption of *serial independence* or *non-autocorrelation*. Algebraically,

$$\begin{aligned} Cov(\varepsilon_i, \varepsilon_j) &= E[\{\varepsilon_i - E(\varepsilon_i)\}\{\varepsilon_j - E(\varepsilon_j)\}] \\ &= E(\varepsilon_i \varepsilon_j) \\ &= 0 \quad \text{for } i \neq j \end{aligned}$$

This assumption implies that ε is independent such that different values of ε are not correlated. The violation of this assumption creates the problem of *serial correlation* or *autocorrelation*.

(iv) *ε is normally distributed.* This assumption is necessary for conducting statistical tests of significance of the parameters estimated. It can be justified by invoking

[5] For the meaning of bias of an estimate, refer to Chapter 2, Section 2.5.

the Central Limit Theorem. We suppose that ε captures the impact of all omitted variables, and we have many such variables which are minor but are independently distributed random variables. Then, the distribution of the sum of such independently distributed random variables tends to be normal as the number of such variables increases independently. It is obvious that violation of this assumption will render the usual tests of significance (e.g., the *t*-test) for the estimated coefficients inapplicable.

(v) *X is a non-stochastic variable with fixed values in repeated samples.* This implies that the values of *X* are either controllable or fully predictable. Violation of this assumption creates problems like *errors in variables* and *autoregression.*

Probability Distribution of Y_i

Given the above assumptions, we may now look at the probability distribution of the dependent variable Y_i. As Y_i is a linear function of ε_i which is normally distributed, it follows that Y_i is also normally distributed. Further, the mean and variance of Y_i are

$$\begin{aligned} E(Y_i) &= E(\alpha+\beta X_i+\varepsilon_i) \\ &= E(\alpha+\beta X_i)+E(\varepsilon_i) \\ &= \alpha+\beta X_i \quad [\because E(\varepsilon_i)=0 \text{ and the concept of expectation does not relate to } \alpha, \beta, \text{ and } X_i] \end{aligned}$$

$$\begin{aligned} Var(Y_i) &= E[Y_i - E(Y_i)]^2 \\ &= E(\varepsilon_i^2) \qquad (\because Y_i = \alpha+\beta X_i+\varepsilon_i) \\ &= \sigma^2 \end{aligned}$$

So we can write the probability distribution of Y_i as

$$Y_i \sim N(\alpha+\beta X_i, \sigma^2)$$

1.5. DATA FOR ECONOMETRIC ANALYSIS

For empirical analysis of economic problems using the tools of econometrics, we use various types of data. While some econometric tools can straightaway be applied to all types of data, we may require specialized tools for analysing data sets having special features. The three types of data used are cross-sectional data, time series data, and panel data.

Cross-Sectional Data

Cross-sectional data, also known as micro-data, are those collected for different entities in a single time period. Thus, a cross-sectional data set may consist of a sample of individuals, households, firms, regions, countries, or any other type of units at a specific time point. The cross-sectional variables are usually denoted by subscript *i* with $i = 1, 2, \ldots, N$, where

N is the number of cross-sectional units from which data have been collected. Among the important fields where cross-sectional data are extensively used are agricultural economics, industrial economics, labour economics, health economics, urban economics, demography, etc. Recently, a new branch of econometrics has emerged that deals with tools specifically required for analysing cross-sectional data or micro-data. This is called *microeconometrics.*

Time Series Data

Time series data, also called macro-data, are those that are collected for the same entity for different time periods. Thus, a time series data set consists of observations on one or more variables over time. The examples of time series data are GDP, money supply, exports, imports, government expenditure, exchange rates, stock prices, etc. The issue of *data frequency* is important in the context of time series data. For economic variables, the most common data frequencies are annual, quarterly, monthly, weekly, and daily. The time series data are denoted by subscript t with $t = 1, 2, \ldots, T$, where T is the number of time points for which data have been collected. The branch of econometrics that deals with specialized tools for analysis of time series data is popularly known as *macroeconometrics* or *time series econometrics.*[6]

Panel Data

Panel data, also called *longitudinal data,* are data collected for multiple entities where each entity is observed in two or more time points. For instance, if we collect data on some macroeconomic variables (GDP, money supply, exports, etc.) for some countries for two or more years, and arrange these data in a systematic manner, then our data set is called the panel data set. Thus, the panel data set has both cross-sectional and time series dimensions. The panel data are denoted by both i and t subscripts that we used earlier for cross-sectional and time series data, respectively. If the GDP data have been collected from N number of countries and for each country data have been recorded for T number of years, then the GDP variable (labelled as GDP_{it}) would have NT observations. Apart from having enhanced number of observations, the panel data are found to be useful to tackle many econometric problems and analysing specific issues which are otherwise difficult to understand using only cross-sectional or time series data. That is why panel data have been gaining wide applications in recent years in many econometric analyses so that another new branch has emerged which is known as *panel data econometrics.*

Experimental and Non-experimental Data

While discussing the types of data, it is important to note that in econometrics, we use non-experimental data. In fact, econometrics evolved as a discipline separate from mathematical statistics to analyse the non-experimental data. While the experimental data generated in laboratory environments are used in the natural sciences, the econometrician uses non-experimental data (e.g., data collected through sample surveys) that are generated not

[6] Some scholars described *microeconometrics* and *macroeconometrics* as the twin sisters in econometrics.

through controlled or laboratory experiments. Of course, the econometrician draws upon the tools from mathematical statistics whenever necessary, but uses separate (econometric) methods developed to analyse the non-experimental data.

Sources of Data

The data to be used for econometric analysis may be obtained from published sources or collected through field surveys. The former is called *secondary* data and the latter as *primary* or field data. Secondary data are mostly available from government departments and international organizations, such as International Monetary Fund (IMF), World Bank, United Nations Development Programme (UNDP), Food and Agricultural Organization (FAO), and International Food Policy Research Institute (IFPRI). These organizations publish both cross-sectional and time series data which are extensively used by the researchers especially for inter-country comparison. In India, important sources of secondary data are the Reserve Bank of India (RBI), Central Statistical Organisation (CSO), Census of India, National Sample Survey Organisation (NSSO), Planning Commission, and different central and state government ministries. Nowadays, a lot of secondary data on diverse aspects of the Indian economy are collated and sold by many private organizations [e.g., Centre for Monitoring Indian Economy (CMIE), Indiastat.com, EPW Research Foundation]. To analyse different developmental issues, the researchers in India depend heavily on data available from these sources. However, it is to be remembered that data available from the secondary sources should not be utilized without criticism. In particular, one should look into the methodology of data collection and ascertain the quality/reliability of data before they are put to final use.

While secondary data are already published, which the researcher merely gathers as per her/his requirements, quite often such data are found inadequate to analyse the problems under consideration. So the researcher might like to collect more data through field surveys which are called primary data. Here again several methodological issues, such as questionnaire design, sample selection, determination of sample size, time frame for survey, and so on, need to be settled before going to the field for actual collection of data.[7] Needless to mention, data collected following scientifically designed sampling procedure would be more reliable and suitable for the purpose of econometric analysis.

1.6. ABOUT EVIEWS SOFTWARE PACKAGE

To illustrate application of various tools of econometrics discussed in this book, we have used the EViews software package. The full form of EViews is 'Econometric Views'. It is developed by Quantitative Micro Software (QMS), USA. EViews version 1.0 was launched in March 1994 replacing its predecessor, MicroTSP. The latest version is EViews 8, released in March 2013. The exercises carried out in this book are based on this latest version. However, as the dialogues and commands for version 8 are almost the same as the previous two versions

[7] For detailed discussions on these issues, one may refer to Desai and Porter (2006); Devereux and Hoddinott (1993); and Rudra (1989).

(EViews 6 and 7), scholars using the previous version will not find any difficulty in following the illustrative examples contained in this book.

An important feature of EViews is that it is a very simple and user-friendly econometric software package that runs on Windows machines.[8] It takes advantage of the visual features of the modern Windows software so that one can just use the mouse to guide the operations with standard Windows menus and dialogues. For input and output, it supports various formats, including Excel, PSPP/SPSS, DAP/SAS, Stata, RATS, and TSP. Although EViews is specially designed to work with time series data, it is also extensively used for econometric analyses of cross-sectional and panel data. So it is a very versatile software package.

Some important basic capabilities of EViews 8 are as follows.

- Reading and writing of data files in standard spreadsheet formats
- Computing a new series, based on a formula of any complexity
- Obtaining plots of data series, scatter diagrams, bar graphs, pie charts, etc.
- Descriptive statistics: correlations, covariances, autocorrelations, cross-correlations, and histograms
- Ordinary least squares regression, least squares with autoregressive correction and two-stage least squares
- Non-linear least squares
- Robust least squares
- Tests for various econometric problems: heteroskedasticity, autocorrelation, multicollinearity, model specification error, etc.
- Estimating quintile regression
- Breakpoint testing (both for single and multiple breakpoints)
- Breakpoint regression (automatic selection and user-specified)
- Probit and logit estimation of binary choice models
- Linear and non-linear estimation of systems of equations
- Pooled cross-sectional–time series estimation and forecasting
- Various tests of stationarity of time series data
- Causality tests including panel causality testing
- ARCH-GARCH estimation and forecasting
- Estimation and analysis of vector autoregressive (VAR) systems
- Bayesian VARs
- State space models and Kalman filter
- Forecasts based on regression

[8] Agung (2011, xv), who worked with several software packages and has written useful books on software applications, observed that not only is EViews the most user-friendly among the available statistical/econometric software packages but is also equally competent to analyse all kinds of data sets—cross-sectional, panel, and time series. It may also be noted that the EViews software package is quite economical in relation to other statistical/econometric software packages. For instance, one unit of the student version of EViews 8 is currently available for $39.95 only. This software is marketed by IHS Global Inc., USA.

2. CLASSICAL LINEAR REGRESSION MODEL (CLRM)
3. FURTHER DEVELOPMENT AND ANALYSIS OF THE CLRM

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:

$$y = a + bx$$

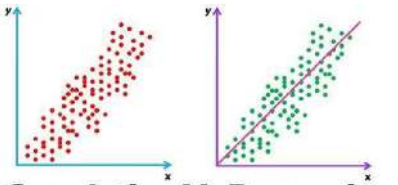
Where, a = constant and b = regression coefficient. In this equation, a and b are two regression parameters.

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 Y_i is dependent variable, X_i is the independent or explanatory variable, α is the regression constant or intercept, β is the regression coefficient for the effect of X_i on Y_i or slope of the regression equation, and e_i is the error we make in predicting Y_i from X_i.

$$Y_i = \alpha + \beta X_i + e_i$$

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.

For details discussion on Classical Linear Regression Model, one may follow the chapter of Simple Linear Regression Model as follows.

[Important: Point 2.6, 2.10, 2.11 and 2.12 are not required for beginners]

2 The Simple Linear Regression Model

This chapter begins with discussion on regression technique, the main workhorse in the toolkit of any researcher, especially in the social sciences. We discuss the issues connected with the two-variable or simple linear regression model (SLRM) in this chapter. The issues specifically dealt with are specification and assumptions of the SLRM, estimation of such a model, hypothesis testing or inference, measuring the quality or goodness of fit of the estimated model, and so on. We also describe the steps involved in estimation of the SLRM using the EViews software package, and clarify the interpretation of EViews regression output.

2.1 DEFINITION

Regression analysis is one of the most important tools at the disposal of any researcher.[1] In very general terms, regression is concerned with describing and evaluating the functional/causal relationship among variables.[2] For instance, if we are interested to know the relationship between consumption expenditure of the individuals and their levels of income, education, sex, caste, and so on, we can do so in terms of regression analysis. In regression analysis, we begin by expressing the relationship among the variables in the form of an equation or a model. In such a model, one of the variables is taken as the dependent variable[3] and all other variables are independent variables.[4] If we are interested to study the relationship between two variables only, our model involves one dependent variable and one independent variable. Such a model is called the simple regression model. On the other hand, we examine the relationship between dependent variable and more than one explanatory variable together in a multiple regression model.

The main issues that concern a researcher employing regression technique to study the relationship among variables are specification and assumptions of the regression model, estimation of such a model, hypothesis testing or inference, and measuring the quality or

[1] The term 'regression' was coined by Francis Galton, an accomplished scientist of nineteenth century and a cousin of Charles Darwin.

[2] While regression shows causal/functional relationship between the variables, correlation shows degree of association between them.

[3] Alternative names of dependent variable are: explained variable, response variable, and regressand.

[4] Independent variable is also called explanatory variable, control variable, and regressor.

goodness of fit of the estimated model. In this chapter, we discuss these and a few other issues in the context of simple regression model while the next chapter discusses these issues in the context of multiple regression model.

2.2 SPECIFICATION AND ASSUMPTIONS

Let us suppose that consumption expenditure depends only on income. A sensible first step to test if at all there is any relation between consumption expenditure and income would be to collect some data on these two variables and display such data in a scatter diagram. Table 2.1 presents data on two variables—annual per capita consumption expenditure and annual per capita NSDP (net state domestic product—which proxies for income)—for 22 major states of India. When we plot these data in a scatter diagram, it appears that there is an approximate linear and positive relationship between per capita consumption expenditure (Y_i) and per capita NSDP (X_i) (see Figure 2.1). This means that increases in Y_i are usually accompanied by increases in X_i, and such a relation between the two variables can be described approximately by a straight line. Then the question is: how to draw such a line? Of course, we could draw such a line by hand and see the intercept and slope of it to form an understanding about the relation between consumption and income. However, in practice such a method is likely to be inaccurate and laborious. Therefore, it would be of interest to determine the extent to which the relationship between Y_i and X_i can be described by an algebraic equation that can be estimated using a defined procedure. Here we may consider using the following equation that 'best fits' our data.

$$Y_i = \alpha + \beta X_i \tag{2.1}$$

However, the problem is that equation (2.1) is an 'exact' one which means, given the values of the intercept (α) and slope (β), we can determine with certainty the value of Y for each value of X. But our scatter diagram suggests that for each value of X, there could be different values of Y. In other words, the relationship between Y and X is not exact, and hence the above model appears inadequate to describe the relationship between Y and X as suggested by the data. In this situation, to make the model realistic, we add a stochastic disturbance term or error term (ε_i) to our model and write it as

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{2.2}$$

A model such as (2.2) is called the two-variable or simple linear regression model (SLRM).[5] As noted in the previous chapter, the stochastic disturbance term ε_i serves several functions. Most notably, it captures the effects of factors other than income on consumption expenditure, corrects for errors in the measurement of Y_i, captures the effects of random and unpredictable events on Y_i, and so on.

[5] The linearity of this model implies that (a) one-unit change in X has the same effect on Y irrespective of the initial value of X (this is so because $\Delta Y_i = \beta \Delta X_i$ when $\Delta \varepsilon_i = 0$) and (b) the highest power of the two parameters α and β is 1.

Table 2.1 Annual Per Capita Consumption Expenditure (PCEXP) and Annual Per Capita NSDP (PCNSDP) in Major States of India

State	PCEXP	PCNSDP
Andhra Pradesh	18,166	51,025
Assam	14,088	27,197
Bihar	10,162	16,119
Chhattisgarh	12,708	38,059
Gujarat	16,093	63,961
Haryana	19,958	78,781
Himachal Pradesh	21,272	50,365
Jammu & Kashmir	15,833	30,582
Jharkhand	12,542	30,719
Karnataka	16,530	50,676
Kerala	21,664	59,179
Madhya Pradesh	13,454	27,250
Maharashtra	20,090	74,027
Orissa	12,854	33,226
Punjab	19,330	62,153
Rajasthan	14,538	34,189
Sikkim	17,653	48,937
Tamil Nadu	16,969	62,499
Tripura	15,595	35,799
Uttar Pradesh	12,778	23,132
Uttarakhand	17,458	55,877
West Bengal	15,424	41,469

Sources: PCEXP (in rupees) for year 2009–10 from NSSO (2011) and PCNSDP (in rupees) for 2009–10 from RBI (2011).

Having specified the relation between Y and X as above, our next task is to estimate the unknown parameters α and β of the model. There are various methods of estimation such as ordinary least squares (OLS) method, maximum likelihood method, moment method, etc. However, we focus on the OLS method as it is the most popular method to estimate a linear model.

The OLS method presupposes fulfilment of the following set of assumptions (which are also referred to as the Gauss-Markov assumptions).

(i) Zero mean of disturbances (ε_i): Using notations, $E(\varepsilon_i) = 0$ for all i.
(ii) Homoskedasticity or constant variance of ε_i: $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ = constant for all i.
(iii) Serial independence of ε_i: $Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.
(iv) Non-stochasticity of X_i: This means the series for X_i is fixed in repeated samples.

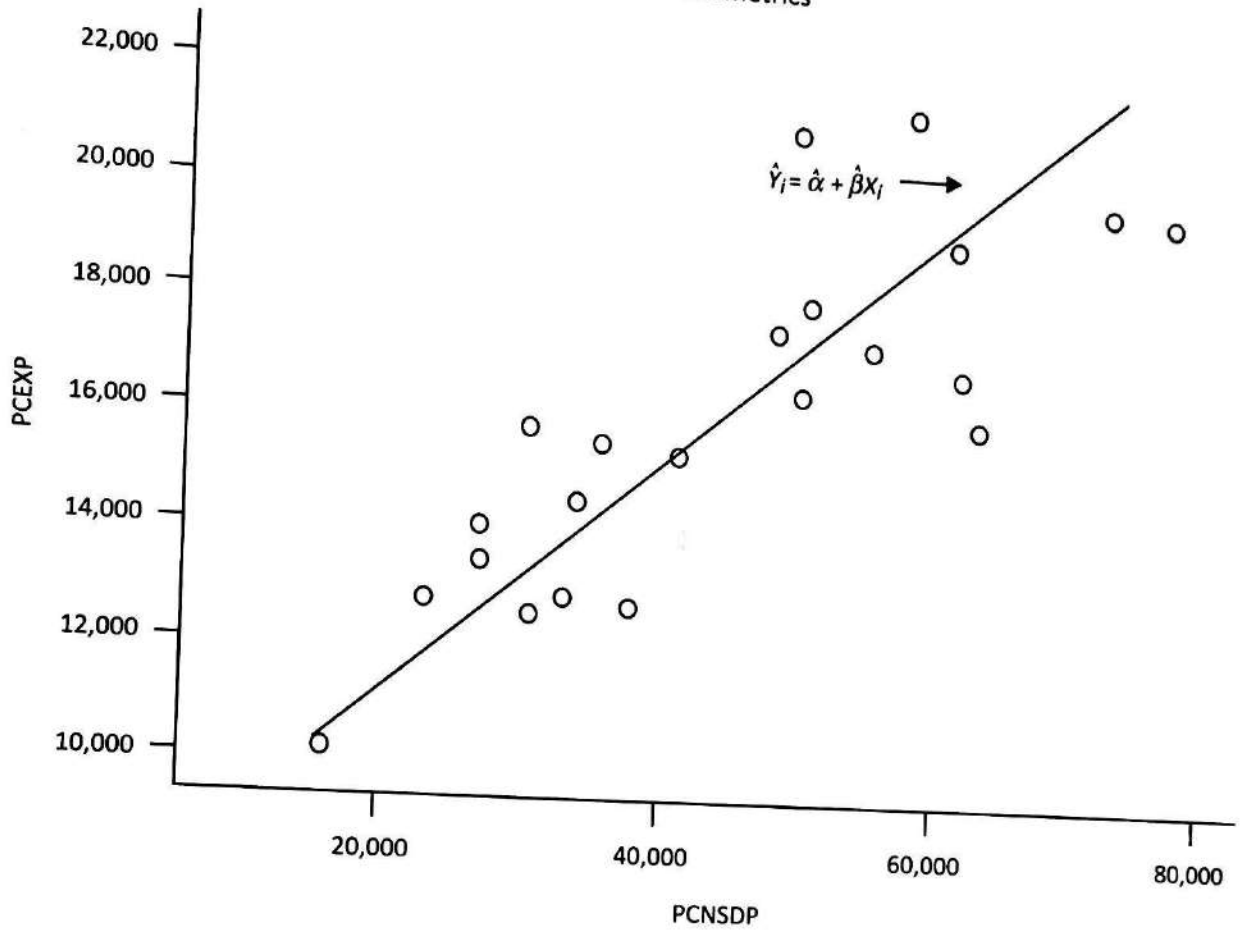


Figure 2.1 Scatter Plot of Data on PCEXP and PCNSDP in Indian States

Source: Author's own.

(v) Exogeneity: This means that the disturbance term ε_i and the explanatory variable X_i are independently distributed, i.e., they are not correlated with each other.

(vi) Linearity: The model must be linear in parameters because the OLS is a linear estimation technique.

(vii) The number of observations (n) is greater than number of parameters in the model, i.e., $n > 2$.

(viii) Normality: Apart from above assumptions, we make the normality assumption for the disturbance term, ε_i, which is actually required to conduct tests of hypotheses after OLS estimation of the model.

2.3 OLS ESTIMATION

Before discussing the OLS method for estimating the above model (2.2), let us note the difference between the terms 'estimator' and 'estimate'. An 'estimator' is a formula or method to estimate some unknown parameter of the model. The 'estimate' is the numerical value obtained after applying such a formula to a given data set.

The model such as (2.2) is called the population regression model as it refers to the relationship between consumption and income with reference to the population data. Here α and β are called 'true' population parameters. Our objective is to obtain numerical values

(which are called estimates) for these unknown parameters by using sample data on Y_i and X_i. Suppose the numerical estimates of α and β are $\hat{\alpha}$ and $\hat{\beta}$, respectively. Using these estimates, we can write our estimated regression line as

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \tag{2.3}$$

Here the series $\hat{Y}_i$ gives the estimated or predicted values of Y_i for different values of X_i, given the values of $\hat{\alpha}$ and $\hat{\beta}$. So given the data on Y_i and X_i, our objective is to compute the values of $\hat{\alpha}$ and $\hat{\beta}$. Once this is done, we say that estimation is complete and the estimated line has been determined. Figure 2.1 shows one such estimated line.

Now the question is to what extent our estimated line is appropriate to describe the relation between consumption and income that emerges from our sample data set? This is an important question because whichever line is considered, some data points will lie above the line and some below the line. In other words, we have some residuals (e_i) from the line, which are defined as

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i \tag{2.4}$$

Here, $i = 1, 2, \ldots, n$ so that we have n sample residuals. In this situation, we may think of choosing the estimated line (i.e., choosing the values of $\hat{\alpha}$ and $\hat{\beta}$) in such a way that the residuals are small. One possible criterion here is to select $\hat{\alpha}$ and $\hat{\beta}$ to make $\sum e_i = 0$. This implies:

$$\sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \tag{2.5}$$

which, on dividing through n, gives

$$\overline{Y} = \hat{\alpha} + \hat{\beta} \overline{X} \tag{2.6}$$

Equation (2.6) implies that $\hat{\alpha}$ and $\hat{\beta}$ should be chosen in such a way that the estimated line passes through $(\overline{X}, \overline{Y})$. However, the problem of choosing an appropriate estimated line is not solved yet. This is because we could pass a line with any slope whatsoever through $(\overline{X}, \overline{Y})$ which would satisfy the condition that the algebraic sum of the residuals is zero (i.e., $\sum e_i = 0$). Therefore, this criterion is inadequate to determine a specific estimated line.

In this situation, we apply the *least-squares criterion* which requires that the values of $\hat{\alpha}$ and $\hat{\beta}$ are chosen in such a way that $\sum e_i^2$ is minimized.[6]

[6] There is a controversy as to who first developed the method of least squares to fit a line. One view is that in the late 1700s and early 1800s Carl Friedrich Gauss in Germany, Adrien-Marie Legendre in France, and Robert Adrain in the United States independently developed the method of least squares. However, Stigler (1986) opined that Gauss probably possessed the method well before others, but he failed to communicate it to his contemporaries.

Using (2.4), we can write:

$$\sum e_i^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \tag{2.7}$$

The necessary conditions of minimization of $\sum e_i^2$ are

$$\frac{\partial \sum e_i^2}{\partial \hat{\alpha}} = 0 \text{ and } \frac{\partial \sum e_i^2}{\partial \hat{\beta}} = 0$$

Applying these conditions, we obtain

$$\sum Y_i = n\hat{\alpha} + \hat{\beta} \sum X_i \tag{2.8a}$$

$$\sum X_i Y_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2 \tag{2.8b}$$

These are termed as the OLS 'normal equations'.

It follows that given data on Y_i and X_i to estimate the line implied by equations 2.8a and 2.8b, which is called the estimated regression line, we have to compute five quantities from the sample data, which are

$$n, \sum X_i, \sum Y_i, \sum X_i Y_i, \text{ and } \sum X_i^2$$

Substitution of these into equations (2.8a) and (2.8b) gives two simultaneous equations which can be solved for the two unknowns, $\hat{\alpha}$ and $\hat{\beta}$. Using the values of $\hat{\alpha}$ and $\hat{\beta}$ in (2.3) we obtain the estimated regression line as

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

2.4 PROPERTIES OF OLS REGRESSION LINE

Some important properties associated with the OLS regression line are as follows:

(i) As previously mentioned, the *OLS regression line passes through the point of means* $(\overline{X}, \overline{Y})$. This follows directly from the first normal equation (2.8a), which, on dividing by n, gives

$$\overline{Y} = \hat{\alpha} + \hat{\beta} \overline{X}$$

(ii) *The residuals e_i have zero covariance with the sample X_i values, and also with $\hat{Y}_i$, which represents the estimated or predicted values of Y_i.* These can be proved as follows.

By definition,

$$\begin{aligned} Cov(X_i, e_i) &= \frac{1}{n}\sum(X_i - \overline{X})(e_i - \overline{e}) \\ &= \frac{1}{n}\sum(X_i - \overline{X})e_i \quad (\because \overline{e} = 0) \\ &= \frac{1}{n}\sum X_i e_i - \frac{1}{n}\overline{X}\sum e_i \\ &= \frac{1}{n}\sum X_i e_i \qquad (\because \sum e_i = 0) \end{aligned}$$

The condition $\frac{\partial \sum e_i^2}{\partial \hat{\beta}} = 0$ implies

$$\begin{aligned} &\frac{\partial}{\partial \hat{\beta}}\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = 0 \\ &\Rightarrow -2\sum X_i(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0 \\ &\Rightarrow -2\sum X_i e_i = 0 \quad [\because Y_i - \hat{\alpha} - \hat{\beta}X_i = Y_i - (\hat{\alpha} + \hat{\beta}X_i) = Y_i - \hat{Y}_i = e_i] \\ &\Rightarrow \sum X_i e_i = 0 \end{aligned}$$

Thus,

$$Cov(X_i, e_i) = \frac{1}{n}\sum X_i e_i = 0$$

Again, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ implies that $\hat{Y}_i$ is a linear function of X_i so that

$$Cov(\hat{Y}_i, e_i) = 0$$

(iii) *The estimated coefficients* $\hat{\alpha}$ and $\hat{\beta}$ *may also be computed sequentially using the following formulae.*

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{2.9a}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} \tag{2.9b}$$

where

$$x_i = X_i - \overline{X}$$

and

$$y_i = Y_i - \overline{Y}$$

Note that (2.9b) is a mere rearrangement of equation (2.8a). Equation (2.9a) follows from substitution of equation (2.9b) into (2.8b). This is shown below.

$$\begin{aligned}
&\Sigma X_i Y_i = (\bar{Y} - \hat{\beta}\bar{X})\Sigma X_i + \hat{\beta}\Sigma X_i^2 \\
&\Rightarrow \Sigma X_i Y_i = \bar{Y}\Sigma X_i - \hat{\beta}\bar{X}\Sigma X_i + \hat{\beta}\Sigma X_i^2 \\
&\Rightarrow \hat{\beta}\left[\Sigma X_i^2 - \bar{X}\Sigma X_i\right] = \Sigma X_i Y_i - \bar{Y}\Sigma X_i \\
&\Rightarrow \hat{\beta}\left[\Sigma X_i^2 - \frac{1}{n}(\Sigma X_i)^2\right] = \Sigma X_i Y_i - \frac{1}{n}\Sigma X_i \Sigma Y_i \\
&\Rightarrow \hat{\beta}\Sigma x_i^2 = \Sigma x_i y_i \\
&\Rightarrow \hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}
\end{aligned}$$

Alternatively, using the definitions of $Cov(X_i, Y_i)$ and $Var(X_i)$, we can write

$$\hat{\beta} = \frac{\Sigma x_i y_i / n}{\Sigma x_i^2 / n} = \frac{\left[\Sigma(X_i - \bar{X})(Y_i - \bar{Y})\right]/n}{\Sigma(X_i - \bar{X})^2 / n} = \frac{Cov(X_i, Y_i)}{Var(X_i)} \tag{2.10}$$

(iv) *The total variation in Y_i may be expressed as sum of two components—the variation 'explained' by the estimated regression line and the variation not explained or 'unexplained' by the estimated regression line.*

To illustrate this properly, consider Figure 2.2, where we have drawn the estimated regression line $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$. Since the least squares regression line passes through the point of means, we take $(\bar{X}, \bar{Y})$ as the new origin. Now consider point P with co-ordinates (X_i, Y_i).

The first co-ordinate can also be expressed as x_i, while the second co-ordinate can be represented as y_i. Now, as shown in the figure, y_i can be split into two components so that

$$y_i = \hat{y}_i + e_i \tag{2.11}$$

where

$$\hat{y}_i = \hat{Y}_i - \bar{Y}.$$

Now squaring and summing over all observations,

$$\begin{aligned}
&\Sigma y_i^2 = \Sigma(\hat{y}_i + e_i)^2 \\
&= \Sigma \hat{y}_i^2 + \Sigma e_i^2 + 2\Sigma \hat{y}_i e_i \\
&= \Sigma \hat{y}_i^2 + \Sigma e_i^2 \qquad [\because Cov(\hat{y}_i, e_i) = 0]
\end{aligned}$$

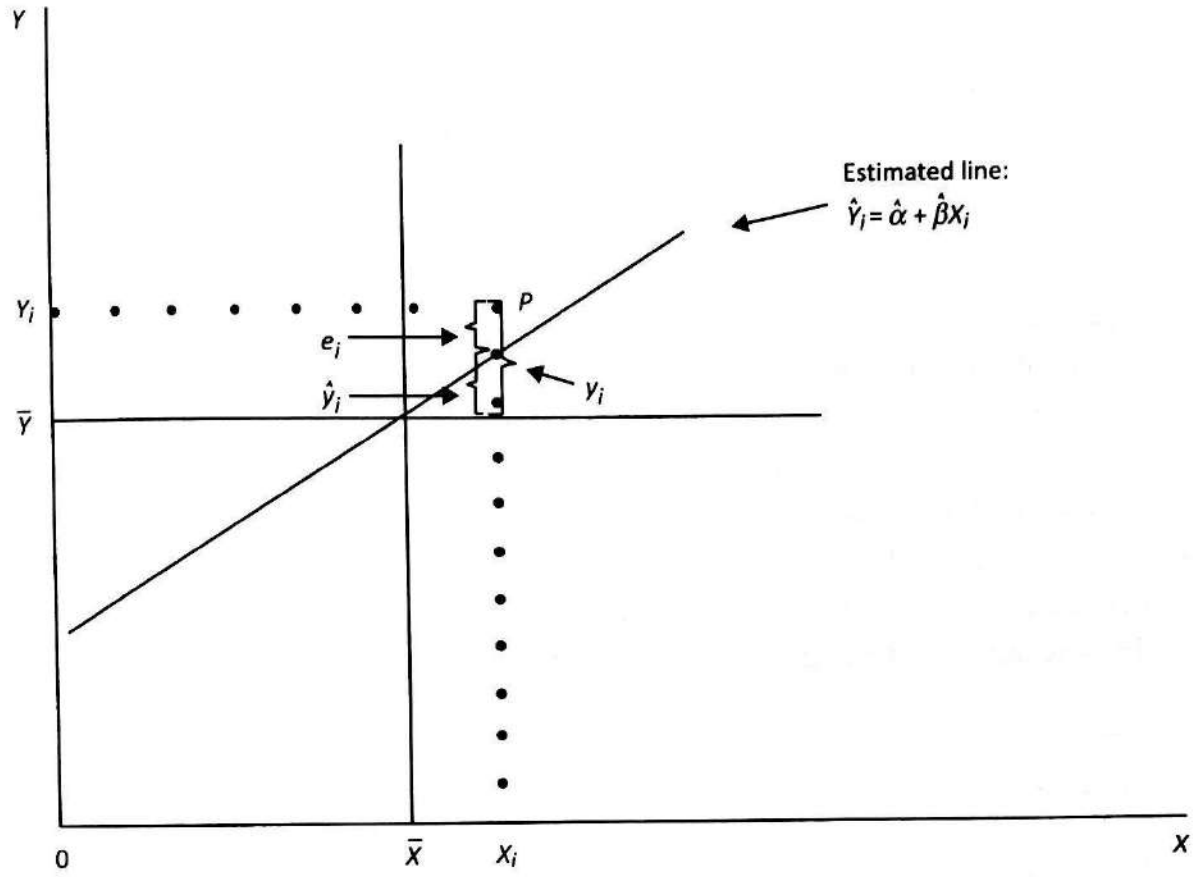


Figure 2.2 Decomposition of Total Sum Squares

Source: Author's own.

We call $\sum y_i^2$ as total sum squares (TSS), $\sum \hat{y}_i^2$ as explained sum squares (ESS), and $\sum e_i^2$ as residual sum squares (RSS). Thus,

$$TSS = ESS + RSS$$

From this result, we may also derive an alternative expression to compute ESS.

$$\begin{aligned} ESS &= TSS - RSS \\ &= \sum y_i^2 - \sum e_i^2 \\ &= \sum \hat{y}_i^2 \\ &= \sum (\hat{\beta} x_i)^2 \\ &= \hat{\beta}^2 \sum x_i^2 \\ &= \left(\frac{\sum x_i y_i}{\sum x_i^2} \right)^2 \sum x_i^2 \\ &= \hat{\beta} \sum x_i y_i \end{aligned}$$

2.5 PROPERTIES OF ESTIMATORS

To choose between estimating principles, we look into the properties satisfied by them. These properties are classified into two groups—small sample properties and large sample properties.[7]

Small Sample Properties

- *Unbiasedness*: An estimator $\hat{\beta}$ is said to be an unbiased estimator of β if its mean or expected value is equal to the value of true population parameter, β, i.e., $E(\hat{\beta}) = \beta$.[8] In everyday terms, this means that if repeated samples of a given size are drawn, and $\hat{\beta}$ computed for each sample, the average of such $\hat{\beta}$ values would be equal to β. However, if $E(\hat{\beta}) \neq \beta$ or $E(\hat{\beta}) - \beta \neq 0$, then $\hat{\beta}$ is said to be biased and the extent of bias for $\hat{\beta}$ is measured by $E(\hat{\beta}) - \beta$.
- *Minimum Variance or Bestness*: An estimator $\hat{\beta}$ is said to be a minimum variance or best estimator of β if its variance is less than the variance of any other estimator, say β^*. Thus, when $Var(\hat{\beta}) < Var(\beta^*)$, $\hat{\beta}$ is called the minimum variance or best estimator of β.[9]
- *Efficiency*: $\hat{\beta}$ is an efficient estimator if the following two conditions are satisfied together:

  (i) $\hat{\beta}$ is unbiased and
  (ii) $Var(\hat{\beta}) \leq Var(\beta^*)$

  An efficient estimator is also called as a minimum variance unbiased estimator (MVUE) or best-unbiased estimator.

- *Linearity*: An estimator is said to have the property of linearity if it is possible to express it as a linear combination of sample observations.[10]
- *Mean-Squared Error* (MSE): Sometimes a difficult choice problem arises while comparing two estimators. Suppose we have two different estimators of which one has lower bias, but higher variance, compared with the other. In other words, when

$$(bias\hat{\beta}) > (bias\beta^*)$$

[7] There is no hard and fast rule to distinguish between small and large samples. However, a working definition is that a small sample has 30 or less observations while the large sample has more than 30 observations.

[8] The unbiasedness property reflects on the *accuracy* of the estimator. Thus, when an estimator is unbiased, we say that it is able to provide an accurate estimate of its true population parameter.

[9] Minimum variance is associated with *reliability* dimension of the estimator. Obviously, the estimator that has lower variance is more reliable than the estimator which has higher variance.

[10] Put differently, linearity is associated with linear (i.e., additive) calculation rather than multiplicative or non-linear calculation.

but

$$Var(\hat{\beta}) < Var(\beta^*)$$

then how to choose between the two estimators? In this situation, where one estimator has a larger bias but a smaller variance than the other estimator, it is intuitively plausible to consider a trade-off between the two characteristics. This notion is given a precise, formal, expression in the mean-squared error.

The mean-squared error for $\hat{\beta}$ is defined as

$$\begin{aligned}
&MSE(\hat{\beta}) = E[\hat{\beta} - \beta]^2 \\
&= E[\{\hat{\beta} - E(\hat{\beta})\} + \{E(\hat{\beta}) - \beta\}]^2 \\
&= E\{\hat{\beta} - E(\hat{\beta})\}^2 + E\{E(\hat{\beta}) - \beta\}^2 + 2E[\{\hat{\beta} - E(\hat{\beta})\}\{E(\hat{\beta}) - \beta\}] \\
&= Var(\hat{\beta}) + (bias\hat{\beta})^2
\end{aligned}$$

This is so because the cross-product term vanishes, as shown below.

$$\begin{aligned}
&E[\{\hat{\beta} - E(\hat{\beta})\}\{E(\hat{\beta}) - \beta\}] = E\{\hat{\beta}E(\hat{\beta}) - \hat{\beta}\beta - E(\hat{\beta})E(\hat{\beta}) + E(\hat{\beta})\beta\} \\
&= E(\hat{\beta})E(\hat{\beta}) - E(\hat{\beta})\beta - E(\hat{\beta})E(\hat{\beta}) + E(\hat{\beta})\beta \\
&= 0
\end{aligned}$$

Now, according to the mean-squared error property, if $MSE(\hat{\beta}) < MSE(\beta^*)$, we say that $\hat{\beta}$ has lower mean-squared error, and accept it as an estimator of β.

Large Sample or Asymptotic Properties

These properties relate to the distribution of an estimator when the sample size is large, and approaches to infinity. The important properties here are the following.

- *Asymptotic Unbiasedness*: $\hat{\beta}$ is an asymptotically unbiased estimator of β if

$$\lim_{n \to \infty} E(\hat{\beta}) = \beta$$

  This means that the estimator $\hat{\beta}$, which is otherwise biased, becomes unbiased as the sample size approaches infinity. It is to be noted that if an estimator is unbiased, it is also asymptotically unbiased, but the reverse is not necessarily true.
- *Consistency*: Whether or not an estimator is consistent is understood by looking at the behaviour of its bias and variance as the sample size approaches infinity. If the increase in sample size reduces bias (if there were one) and variance of the estimator, and this

continues until both bias and variance become zero, as $n \to \infty$, then the estimator is said to be consistent. Thus, $\hat{\beta}$ is a consistent estimator if

$$\lim_{n\to\infty}[E(\hat{\beta})-\beta]=0$$

and

$$\lim_{n\to\infty} Var(\hat{\beta})=0$$

2.6 PROPERTIES OF OLS ESTIMATORS

The ordinary least-squares estimators are best, linear, and unbiased. In brief, we say that they are BLUE.[11] The BLUE properties for the least-squares estimators are proved below.

Recall the two-variable linear model

$$Y_i=\alpha+\beta X_i+\varepsilon_i$$

where ε_i satisfies the OLS assumptions:

$$\left.\begin{aligned}E(\varepsilon_i)&=0\\E(\varepsilon_i^2)&=\sigma_\varepsilon^2\end{aligned}\right\}\ \text{for all } i$$
$$E(\varepsilon_i\varepsilon_j)=0 \quad \text{for } i\neq j$$

We know that

$$\hat{\beta}=\frac{\sum x_i y_i}{\sum x_i^2}$$

and

$$\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{X}$$

where

$$x_i=X_i-\bar{X}$$

and

$$y_i=Y_i-\bar{Y}$$

[11] BLUE—best, linear and unbiased estimator. The BLUE property of the OLS estimators is also known as the Gauss-Markov theorem.

Unbiasedness of $\hat{\beta}$

$$\begin{aligned}\hat{\beta} &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i (Y_i - \bar{Y})}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i - \bar{Y}\sum x_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum x_i^2} \qquad [\because \sum x_i = \sum (X_i - \bar{X}) = 0] \\ &= \sum w_i Y_i \end{aligned} \tag{2.12}$$

where

$$w_i = \frac{x_i}{\sum x_i^2} \tag{2.13}$$

It follows from (2.13) that

$$\sum w_i = \frac{\sum x_i}{\sum x_i^2} = 0 \tag{2.14}$$

$$\begin{aligned}\sum w_i X_i &= \frac{\sum x_i X_i}{\sum x_i^2} \\ &= \frac{\sum (X_i - \bar{X}) X_i}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum X_i^2 - \bar{X}\sum X_i}{\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}^2} \\ &= \frac{\sum X_i^2 - n\bar{X}^2}{\sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2} \\ &= \frac{\sum X_i^2 - n\bar{X}^2}{\sum X_i^2 - n\bar{X}^2} \\ &= 1 \end{aligned} \tag{2.15a}$$

$$\sum w_i^2 = \frac{\sum x_i^2}{(\sum x_i^2)^2} = \frac{1}{\sum x_i^2} \tag{2.15b}$$

From (2.12), we have

$$\begin{aligned}\hat{\beta} &= \sum w_i Y_i \\ &= \sum w_i(\alpha + \beta X_i + \varepsilon_i) \\ &= \alpha \sum w_i + \beta \sum w_i X_i + \sum w_i \varepsilon_i \\ &= \beta + \sum w_i \varepsilon_i \qquad (\because \sum w_i = 0 \text{ and } \sum w_i X_i = 1)\end{aligned} \tag{2.16}$$

Taking expectations,

$$\begin{aligned}E(\hat{\beta}) &= E(\beta + \sum w_i \varepsilon_i) \\ &= \beta + \sum w_i E(\varepsilon_i) \\ &= \beta \qquad [\because E(\varepsilon_i) = 0]\end{aligned}$$

This proves that $\hat{\beta}$ is unbiased.

Linearity of $\hat{\beta}$

From (2.12),

$$\hat{\beta} = \sum w_i Y_i$$

Since w_i's are a set of fixed values, we may write

$$\hat{\beta} = w_1 Y_1 + w_2 Y_2 + ... + w_n Y_n$$

This shows that $\hat{\beta}$ is a linear combination of sample values of Y_i, the dependent variable. Thus, $\hat{\beta}$ has the property of linearity.

Minimum Variance or Bestness for $\hat{\beta}$

In order to prove minimum variance or bestness property for $\hat{\beta}$, we shall compute the variance of $\hat{\beta}$ and show that it is lower than the variance of some other estimator.

From (2.16), we have

$$\begin{aligned}&\hat{\beta} = \beta + \sum w_i \varepsilon_i \\ &\Rightarrow \hat{\beta} - \beta = \sum w_i \varepsilon_i \\ &\Rightarrow \hat{\beta} - E(\hat{\beta}) = \sum w_i \varepsilon_i \qquad [\because E(\hat{\beta}) = \beta]\end{aligned}$$

Thus,

$$\begin{aligned}Var(\hat{\beta}) &= E[\hat{\beta} - E(\hat{\beta})]^2 \\ &= E\left(\sum w_i \varepsilon_i\right)^2\end{aligned}$$

$$
\begin{aligned}
&= E\left(\sum w_i^2 \varepsilon_i^2 + 2\sum_{i<j} w_i w_j \varepsilon_i \varepsilon_j \right) \\
&= \sum w_i^2 E(\varepsilon_i^2) + 2\sum_{i<j} w_i w_j E(\varepsilon_i \varepsilon_j) \\
&= \sigma^2 \sum w_i^2 \qquad [\because E(\varepsilon_i^2) = \sigma^2 \text{ and } E(\varepsilon_i \varepsilon_j) = 0 \text{ for } i \neq j] \\
&= \frac{\sigma^2}{\sum x_i^2} \qquad \left(\because \sum w_i^2 = \frac{1}{\sum x_i^2} \right)
\end{aligned} \tag{2.17}
$$

Let us now consider some other estimator, say β^*, such that

$$\beta^* = \sum c_i Y_i$$

where c_i $(i = 1, 2, \ldots, n)$ represents a set of weights

Then,

$$
\begin{aligned}
\beta^* &= \sum c_i (\alpha + \beta X_i + \varepsilon_i) \\
&= \alpha \sum c_i + \beta \sum c_i X_i + \sum c_i \varepsilon_i
\end{aligned} \tag{2.18}
$$

Taking expectations,

$$E(\beta^*) = \alpha \sum c_i + \beta \sum c_i X_i \quad [\because E(\varepsilon_i) = 0]$$

It is clear that we require the weights to be such that β^* is an unbiased estimator. This imposes the conditions

$$\sum c_i = 0 \text{ and } \sum c_i X_i = 1 \Rightarrow \sum c_i x_i = 1 \tag{2.19}$$

Let us now compute $Var(\beta^*)$ accepting these conditions. Under these conditions, equation (2.18) reduces to

$$\beta^* = \beta + \sum c_i \varepsilon_i \;\Rightarrow\; (\beta^* - \beta) = (\beta^* - E(\beta^*)) = \sum c_i \varepsilon_i$$

Thus,

$$
\begin{aligned}
&Var(\beta^*) = E[\beta^* - E(\beta^*)]^2 \\
&= E\left(\sum c_i \varepsilon_i \right)^2 \\
&= E\left(\sum c_i^2 \varepsilon_i^2 + 2\sum_{i<j} c_i c_j \varepsilon_i \varepsilon_j \right) \\
&= \sum c_i^2 E(\varepsilon_i^2) + 2\sum_{i<j} c_i c_j E(\varepsilon_i \varepsilon_j) \\
&= \sigma^2 \sum c_i^2 \qquad [\because E(\varepsilon_i^2) = \sigma^2 \text{ and } E(\varepsilon_i \varepsilon_j) = 0 \text{ for } i \neq j]
\end{aligned} \tag{2.20}
$$

To compare $Var(\hat{\beta})$ with $Var(\beta^*)$, consider the expression

$$\begin{aligned} c_i &= w_i + (c_i - w_i) \\ \Rightarrow \sum c_i^2 &= \sum w_i^2 + \sum (c_i - w_i)^2 + 2\sum w_i (c_i - w_i) \end{aligned} \tag{2.21}$$

Note that

$$\begin{aligned} &\sum w_i (c_i - w_i) \\ &= \sum w_i c_i - \sum w_i^2 \\ &= \frac{\sum c_i x_i}{\sum x_i^2} - \frac{1}{\sum x_i^2} \quad \left(\because w_i = \frac{x_i}{\sum x_i^2} \right) \\ &= \frac{1}{\sum x_i^2} - \frac{1}{\sum x_i^2} \quad \left[\because \sum c_i x_i = 1, \text{as shown in (2.19) above} \right] \\ &= 0 \end{aligned}$$

Thus,

$$\begin{aligned} Var(\beta^*) &= \sigma^2 [\sum w_i^2 + \sum (c_i - w_i)^2] \\ &= \frac{\sigma^2}{\sum x_i^2} + \sigma^2 \sum (c_i - w_i)^2 \quad \left(\because \sum w_i^2 = \frac{1}{\sum x_i^2} \right) \\ &= Var(\hat{\beta}) + \sigma^2 \sum (c_i - w_i)^2 \end{aligned}$$

Since $\Sigma(c_i - w_i)^2 > 0$ unless $c_i = w_i$ for all i, $Var(\hat{\beta}) < Var(\beta^*)$, and we conclude that $\hat{\beta}$ is a minimum variance or best estimator.

2.7 STATISTICAL INFERENCE IN SLRM

After estimating the population parameters (α and β) of our regression model, our next task is to examine statistical significance of the estimated coefficients ($\hat{\alpha}$ and $\hat{\beta}$) by applying our knowledge of statistical inference.[12] Examination of statistical significance of the estimated coefficients specifically requires the knowledge about their sampling distributions. In this regard, it may be noted that

$$\hat{\alpha} \sim N\left[\alpha, \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2} \right) \right] \tag{2.22}$$

[12] In layman's language, testing statistical significance of an estimated coefficient (say $\hat{\beta}$) enables us to understand the usefulness of X_i as a predictor of Y_i.

$$\hat{\beta} \sim N\left[\beta, \frac{\sigma^2}{\sum x_i^2}\right] \tag{2.23}$$

Expression (2.22) states that $\hat{\alpha}$ has a normal distribution with mean equal to α and variance $\frac{1}{n}+\frac{\bar{X}^2}{\sum x_i^2}$. Similarly, (2.23) states that $\hat{\beta}$ also follows a normal distribution with mean equal to β and variance $\sigma^2 / \sum x_i^2$. However, these results are useful when the variance of the disturbance term (σ^2) is known. Unfortunately, in practice, σ^2 is not known and has to be estimated as

$$\hat{\sigma}^2 = \frac{RSS}{n-2} = \frac{\sum e_i^2}{n-2} \tag{2.24}$$

where $\hat{\sigma}^2$ is the estimate of σ^2.[13]

Hypothesis Testing

We formalize the object of testing statistical significance of $\hat{\beta}$ (also $\hat{\alpha}$) by stating that we want to test the validity of the null hypothesis (H_N)[14] that the value of true population parameter β is zero against the alternative hypothesis (H_A)[15] that it is different from zero. In the present context, we set our hypotheses as

$H_N: \beta = 0$
$H_A: \beta \neq 0$ (under two-tailed test)

However, if we have any prior knowledge about the sign of β (say positive), then the hypotheses are set as

$H_N: \beta = 0$
$H_A: \beta > 0$ (one-tailed test)

[13] Note that $\sqrt{\hat{\sigma}^2} = \hat{\sigma}$ is called the *standard error of regression*. It provides an estimate of standard deviation of the regression error (or disturbance term) ε_i.

[14] In simple language, null hypothesis is what we are going to test.

[15] The alternative hypothesis represents our conclusion if the experimental test indicates that the null hypothesis is false.

Having set the hypotheses to be tested, our next task is to compute one t-value, which is denoted by t^*. The formula used for computation of t^* is:[16]

$$t^* = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$
$$= \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (\text{under } H_N : \beta = 0) \tag{2.25}$$

where $SE(\hat{\beta})$ is the standard error of $\hat{\beta}$.[17]

Having computed the value of t^* in the manner stated above, we compare it with critical (or theoretical) t-value obtained from the t-table for level of significance[18] $\lambda/2$ (under

[16] The reason behind calculation to t^* in this manner lies in the following: As $\hat{\beta}$ is normally distributed, $(\hat{\beta}-\beta)/\sqrt{Var(\hat{\beta})} \sim N(0, 1)$, which means that $(\hat{\beta}-\beta)/\sqrt{Var(\hat{\beta})}$ has a standard normal distribution. Again, it has been be shown that RSS/σ^2 follows a χ^2 distribution with degrees of freedom $k = n - 2$, where n = number of observations. Now there is a theorem which states if we have two variables (X_1 and X_2) which are independent but $X_1 \sim N(0, 1)$ and $X_2 \sim \chi^2$ with degrees of freedom k, then $X_1/\sqrt{X_2/k} = standardised\ normal/\sqrt{independent\ averaged\ \chi^2}$ follows a t-distribution with degrees of freedom k. Using this theorem, we write

$$t^* = \frac{\hat{\beta} - \beta}{\sqrt{Var(\hat{\beta})}} \Big/ \sqrt{\frac{RSS}{\sigma^2 (n-2)}} \quad \text{with degrees of freedom } k = n - 2$$

Thus,

$$t^* = \frac{\hat{\beta} - \beta}{\sqrt{\sigma^2 / \sum x_i^2}} \Big/ \sqrt{\frac{\hat{\sigma}^2 (n-2)}{\sigma^2 (n-2)}} \qquad [\because Var(\hat{\beta}) = \sigma^2 / \sum x_i^2 \text{ and } RSS = \hat{\sigma}^2 (n-2)]$$
$$= \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}}$$
$$= \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$

Similarly, we can say that $\frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})}$ also has a t-distribution with degrees of freedom $n - 2$.

[17] The standard error of an estimator is nothing but the standard deviation of the sampling distribution of the estimator and the sampling distribution of the estimator is a probability or frequency distribution of the set of values of the estimator obtained from all possible samples of the same size from a given population.

[18] Level of significance is the probability of rejecting the null hypothesis (H_N) when it is actually true, i.e., it is the probability of committing a Type I error.

two-tailed test) and degrees of freedom[19] $n-2$. The following *decision rules* are followed here.

- If $|t^*| > t_{\lambda/2}(n-2)$, i.e., absolute value of computed-t is greater than the value of critical-t at level of significance $\lambda/2$ and degrees of freedom $n-2$, then reject the H_N and conclude that $\hat{\beta}$ is statistically significant at significance level $\lambda/2$ and the regression is meaningful.
- On the other hand, if $|t^*| \leq t_{\lambda/2}(n-2)$, then accept the H_N and conclude that $\hat{\beta}$ is statistically insignificant at significance level $\lambda/2$ and the regression is meaningless.

The same procedure and decision rules apply when we test statistical significance of $\hat{\alpha}$.

One-Tailed Test

The one-tailed test reduces the critical-t value for a given degrees of freedom, which increases the possibility of obtaining significant regression result. Application of the one-tailed test is justifiable when we are certain about the sign (positive/negative) of the slope parameter β. For instance, the Keynesian consumption theory suggests that the marginal propensity of consume (represented by β) is positive so that our hypotheses, while estimating consumption–income relationship, may be written as

$$H_N: \beta = 0$$
$$H_A: \beta > 0$$

In this situation, we apply the one-tailed test procedure in order to examine statistical significance of $\hat{\beta}$. The *decision rules* here are

- If $|t^*| > t_{\lambda}(n-2)$, i.e., absolute value of computed-t is greater than the value of critical-t at level of significance λ and degrees of freedom $n-2$, then reject the H_N and conclude that $\hat{\beta}$ is statistically significant at the level of significance λ.
- On the other hand, if $|t^*| \leq t_{\lambda}(n-2)$, then do not reject the H_N and conclude that $\hat{\beta}$ is statistically insignificant at the level of significance λ.

Likewise, when we are certain that $\beta < 0$ (e.g., in investment-rate of interest model), we may continue with the one-sided test. In this situation, the hypotheses are

$$H_N: \beta = 0$$
$$H_A: \beta < 0$$

and the *decision rules* are the same as mentioned above.

[19] The term degrees of freedom means the total number of observations in the sample (n) less the number of independent (linear) constraints put on them. In other words, it is the number of independent observations out of total n observations. In the context of the two-variable model, the degrees of freedom is $n-2$ (not n); for the k-variable model, it is $n-k$. The general rule is that degrees of freedom = n minus number of parameters estimated.

It needs mention that in practice researchers apply one-sided tests more frequently. This is primarily because, compared with two-tailed tests, the critical-t value for rejecting the H_N is lower for the one-sided test, so it is easier to refute the H_N and establish the relationship between the variables (dependent and explanatory) statistically significantly. However, the one-sided test should not be applied mechanically and should be justified beforehand on the basis of theory, previous experience, or common sense.

Confidence Intervals

There is an alternative way of drawing inferences about our null hypothesis (H_N: $\beta = 0$). This is by constructing a confidence interval (CI). The CIs give the numerical region in which we would have some degree of confidence that our H_N: $\beta = 0$ is true.[20] Thus, the CIs capture the region in which H_N: $\beta = 0$ would not be rejected. The formula for the CI for β is given by

$$\hat{\beta} \pm SE(\hat{\beta}) t_{\lambda/2} \tag{2.26}$$

where $t_{\lambda/2}$ is the critical value of t with $\lambda/2$ level of significance and $n - 2$ degrees of freedom.

The *decision rule* is that if β falls within the CI, we do not reject H_N. On the other hand, we reject the H_N when β falls outside the CI. Alternatively, we may say that when the CI includes the value of 0 (zero), we accept the H_N and conclude that $\hat{\beta}$ is statistically insignificant. However, if 0 doesn't fall within the CI, we reject the H_N and conclude that $\hat{\beta}$ is statistically significant.

The *p*-value Approach

The output from many econometric software packages (including EViews), apart from providing computed-t statistics for the estimated coefficients, also provide the p-values which can be used as an alternative approach in assessing the significance of the estimated coefficients (and hence validity of the H_N). The p-value is the smallest significance level at which the H_N could be rejected, based on the test statistic actually observed.[21] The p-value approach is more informative than the 'choice of significance levels and obtain critical values' approach because one can see exactly the level of significance of the estimated coefficient.[22] For example, if $p = 0.03$, it implies that the H_N would be rejected at 3% level of significance. Similarly, $p = 0.15$ implies that the H_N would be rejected at 15% level of significance. It is customary to reject the H_N and conclude that the estimated coefficient under consideration is statistically significant if the p-value is less or equal to 0.10.

[20] Thus, confidence intervals are *interval estimates* as they provide a range of likely values for the population parameter.

[21] The p-value also represents the probability of committing a Type-I error that occurs when we reject a true H_N.

[22] Regression output from almost all econometrics software packages display p-values alongside the estimated coefficients and their standard errors and computed-t values. It is to be noted that the p-value reported by EViews is computed for a two-tailed test. So when we are interested in a one-tailed test, we will have to look up the critical value ourselves.

2.8 MEASURING GOODNESS OF FIT

In the context of the simple regression model, we measure the goodness of fit of the estimated equation by using the *squared-r* (i.e., r^2) statistic, where r is the value of simple correlation coefficient between Y_i and X_i. r^2, being the ratio of *ESS* to *TSS*, shows the proportion of total variation in the dependent variable which is explained by the independent/explanatory variable of the model. Thus,

$$r^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

This may be proved as follows.

$$\begin{aligned} r &= \frac{Cov(X_i, Y_i)}{\sqrt{Var(X_i)}\sqrt{Var(Y_i)}} \\ &= \frac{\sum x_i y_i}{n S_x S_y} \quad (S_x \text{ and } S_y \text{ are standard deviations of } X_i \text{ and } Y_i \text{ respectively}) \\ &= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \end{aligned}$$

Therefore,

$$\begin{aligned} r^2 &= \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \\ &= \hat{\beta}\frac{\sum x_i y_i}{\sum y_i^2} \\ &= \frac{ESS}{TSS} \qquad (\because \hat{\beta}\sum x_i y_i = ESS \text{ and } \sum y_i^2 = TSS \text{ as noted earlier}) \\ &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum e_i^2}{\sum y_i^2} \end{aligned} \tag{2.27}$$

The value of r^2 always lies between 0 and 1. When the value of r^2 is 0, no part of variation in the dependent variable is explained by the variation in explanatory variable of the model. On the other hand, when the value of r^2 is 1, the entire part of variation in dependent variable is explained by the variation in explanatory variable. However, these are extreme cases. In reality, the value of r^2 is found somewhere between 0 and 1. When it is close to 1, we conclude that we have a good fit estimated line; if it is close to 0, we conclude that the estimated line does not provide a good fit.

2.9 ANALYSIS OF VARIANCE ON OLS REGRESSION

We can also set out the test of significance discussed above in an Analysis of Variance (ANOVA) framework. This is more useful in the context of multiple regressions. As before, our null hypothesis here is H_N: $\beta = 0$. Under this approach, we actually test the significance of the ESS as against RSS from regression. This is done in the manner described below.

We know that

$$TSS = ESS + RSS$$

$$TSS = \sum y_i^2$$

$$ESS = \hat{\beta}^2 \sum x_i^2 = \hat{\beta} \sum x_i y_i$$

$$RSS = \sum e_i^2$$

Then,

$\frac{RSS}{\sigma^2}$ follows a χ^2 distribution with degrees of freedom $n-2$; and

$\frac{ESS}{\sigma^2}$ also follows a χ^2 distribution with degrees of freedom 1.

Now assuming that the H_N is true and that these two χ^2s are independent, their ratio divided by respective degrees of freedom gives an F-statistic, which is

$$F = \frac{\frac{ESS}{\sigma^2} \Big/ 1}{\frac{RSS}{\sigma^2} \Big/ n-2} = \frac{ESS/1}{RSS/(n-2)} = \frac{\hat{\beta} \sum x_i y_i / 1}{\sum e_i^2 / (n-2)}$$

To compute the value of F using this formula, we need some information which are obtained from the following ANOVA table.

After computing the value of F, using the information contained in the ANOVA table, we set the *decision rules* as follows.

(i) If $F^* > F_\lambda (1, n-2)$, then reject H_N: $\beta = 0$; and

(ii) If $F^* \leq F_\lambda (1, n-2)$, then do not reject H_N: $\beta = 0$

Table 2.2 ANOVA Table: To Compute F

Source of Variation	Sum Squares	Degrees of Freedom	Mean Squares
X_i	$ESS = \hat{\beta}^2 \sum x_i^2 = \hat{\beta} \sum x_i y_i$	1	$ESS/1$
Residuals	$RSS = \sum e_i^2$	$n-2$	$RSS/(n-2)$
Total	$TSS = \sum y_i^2$	$n-1$	

Source: Author's own compilation.

2.10 SOME IMPORTANT RELATIONS IN THE CONTEXT OF SLRM

Relation between Regression Slope and Correlation Coefficient

There is a relation between the regression slope ($\hat{\beta}$) and correlation coefficient (r) between X_i and Y_i which is demonstrated as follows.

$$r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \qquad (x_i = X_i - \bar{X} \text{ and } y_i = Y_i - \bar{Y})$$

$$= \frac{\sum x_i y_i}{\sum x_i^2} \frac{\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}}$$

$$= \frac{\sum x_i y_i}{\sum x_i^2} \frac{\sqrt{\sum x_i^2/n}}{\sqrt{\sum y_i^2/n}}$$

$$= \hat{\beta} \frac{S_x}{S_y} \qquad (S_x \text{ and } S_y \text{ are standard deviations of } X_i \text{ and } Y_i \text{ respectively})$$

Thus, the relation between the regression slope and correlation coefficient is given by

$$\hat{\beta} = r \frac{S_y}{S_x} \tag{2.28}$$

Relation between *F*-statistic and r^2

We know that

$$TSS = ESS + RSS$$

$$TSS = \sum y_i^2$$

$$ESS = \hat{\beta} \sum x_i y_i = r^2 \sum y_i^2 \qquad \left(\because r^2 = \hat{\beta} \frac{\sum x_i y_i}{\sum y_i^2} \right)$$

$$RSS = TSS - ESS = \sum y_i^2 - r^2 \sum y_i^2 = (1 - r^2) \sum y_i^2$$

Thus,

$$F = \frac{ESS/1}{RSS/(n-2)} = \frac{r^2 \sum y_i^2 / 1}{(1-r^2)\sum y_i^2/(n-2)} = \frac{(n-2)r^2}{1-r^2} \tag{2.29}$$

The implication of this relation is that if we have computed the value of r^2, we may use it to compute the value of F and skip computations involved in the ANOVA table.

Relation between F and t^2

In the two-variable model,

$$\begin{aligned} F &= \frac{ESS/1}{RSS/(n-2)} \\ &= \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum e_i^2/(n-2)} \\ &= \frac{\sum\left[(\hat{\alpha}+\hat{\beta}X_i)-(\hat{\alpha}+\hat{\beta}\bar{X})\right]^2}{\hat{\sigma}^2} \qquad [\because \hat{\sigma}^2 = \sum e_i^2/(n-2)] \\ &= \frac{\hat{\beta}^2 \sum(X_i - \bar{X})^2}{\hat{\sigma}^2} \\ &= \frac{\hat{\beta}^2}{\hat{\sigma}^2/\sum x_i^2} \qquad (x_i = X_i - \bar{X}) \\ &= \frac{\hat{\beta}^2}{[SE(\hat{\beta})]^2} \\ &= t^2 \end{aligned} \tag{2.30}$$

This result implies that in the simple regression or two-variable model, the F-test and the two-tailed t-test on the slope coefficient both have the null and alternative hypotheses as: $H_N: \beta = 0$ and $H_A: \beta \neq 0$.[23] Thus, these tests lead to the same conclusion here. Hence there is no point in performing both the tests in simple regression analysis. However, it will be clear from our discussion in the next chapter that in the multiple regression analysis, the F and t tests have different roles to perform, and they test different null hypotheses.

[23] It is also to be noted that the critical value of F at any given level of significance here is equal to the square of the critical value of t. We refrain from proving this result as such details are beyond our scope.

Relation between r^2 and t-statistic

We can derive the relation between r^2 and t-statistic using (2.29) and (2.30).

$$t^2 = \frac{(n-2)r^2}{1-r^2}$$
$$\Rightarrow r^2 = \frac{t^2}{t^2+(n-2)} \tag{2.31}$$

2.11 REGRESSION WITHOUT INTERCEPT TERM

Sometimes the two-variable regression model may not have the intercept term. Such a model is called *no-intercept model* or *regression through the origin*. An example of no-intercept model is the one that examines relation between output and variable cost. Other examples of no-intercept model abound in economics. For the no-intercept model, the formula for obtaining the estimate for the slope coefficient would not have 'mean corrections'. The formulas for obtaining estimated variance of the disturbance term, variance of estimated slope and r^2 coefficient are also different. These are given below.

Suppose our no-intercept population regression model is

$$Y_i = \beta X_i + \varepsilon_i \tag{2.32}$$

and the estimated model is

$$\hat{Y}_i = \hat{\beta} X_i \tag{2.33}$$

This is obtained by minimizing $\sum e_i^2 = \sum (Y_i - \hat{\beta} X_i)^2$.

Application of the necessary condition of minimization here yields

$$\frac{d\sum e_i^2}{d\hat{\beta}} = 2\sum (Y_i - \hat{\beta} X_i)(-X_i) = 0$$
$$\Rightarrow \sum (Y_i - \hat{\beta} X_i)(X_i) = 0$$
$$\Rightarrow \hat{\beta} \sum X_i^2 = \sum X_i Y_i$$
$$\Rightarrow \hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} \tag{2.34}$$

This is the formula used for computation of $\hat{\beta}$ in the context of no-intercept model.

The variance of $\hat{\beta}$ for the no-intercept model is obtained as follows.

From (2.34),

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2}$$

$$= \frac{\sum X_i(\beta X_i + \varepsilon_i)}{\sum X_i^2}$$
$$= \beta + \frac{\sum X_i \varepsilon_i}{\sum X_i^2} \tag{2.35}$$

Taking expectations,

$$E(\hat{\beta}) = \beta \Rightarrow \hat{\beta} \text{ is unbiased}$$

Therefore,

$$\begin{aligned} Var(\hat{\beta}) &= E(\hat{\beta} - \beta)^2 \\ &= E\left(\frac{\sum X_i \varepsilon_i}{\sum X_i^2}\right)^2 \\ &= \frac{E(\varepsilon_i^2)}{\sum X_i^2} \\ &= \frac{\sigma^2}{\sum X_i^2} \end{aligned} \tag{2.36}$$

Here σ^2 is unknown but can be estimated as

$$\hat{\sigma}^2 = \frac{RSS}{n-1}$$

Now, the standard error of $\hat{\beta}$ is

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{\sum X_i^2}}$$

Hypothesis Testing

The *t*-test procedure to test H_N: $\beta = 0$ against H_A: $\beta \neq 0$ holds in the context of no-intercept model, but the formula for computation of $SE(\hat{\beta})$ is different. Here computed-*t* (i.e., t^*) is calculated as

$$t^* = \frac{\hat{\beta}}{SE(\hat{\beta})} = \hat{\beta} \Big/ \sqrt{\frac{\hat{\sigma}^2}{\sum X_i^2}} \tag{2.37}$$

The *decision rules* are same as in the intercept-present model.

Goodness of Fit

For the model without the intercept term, the sum of residuals (i.e., Σe_i) do not necessarily add up to zero like the model with intercept term. Further, the 'fundamental identity' (i.e., $TSS = ESS + RSS$) is no longer true in general for the no-intercept model. For this reason, the conventional r^2 formula used for the intercept-present model is not appropriate to understand the goodness of fit of the no-intercept model. Using the conventional r^2 formula for the no-intercept model may produce negative value of r^2 in some cases (when $RSS > TSS$) and we are unable to interpret properly the value of r^2. To rule out such a possibility, it is essential to use an appropriate r^2 formula for the no-intercept model, which is obtained by modifying the expression of the fundamental identity in the following manner.

For the intercept-present model,

$$TSS = ESS + RSS$$

i.e.,

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(\hat{Y}_i - \bar{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2$$

For the no-intercept model, we replace $\bar{Y}$ by zero (as the regression line here passes through the origin) so that the expression for the fundamental identity becomes

$$\Sigma Y_i^2 = \Sigma \hat{Y}_i^2 + \Sigma e_i^2$$

Therefore,

$$r^2 = \frac{\Sigma \hat{Y}_i^2}{\Sigma Y_i^2} = 1 - \frac{\Sigma e_i^2}{\Sigma Y_i^2} \tag{2.38}$$

This is the appropriate formula to compute the value of r^2 when intercept term is absent in the model.[24]

2.12 REVERSE REGRESSION

Until now, we considered the regression of Y_i on X_i. This is called *direct regression*. However, sometimes we may have to consider the regression of X_i on Y_i as well. This is called *reverse regression*. Let us write our reverse regression model as

$$X_i = \alpha^* + \beta^* Y_i + v_i$$

[24] The *r^2-statistic* for the no-intercept model is also called the *raw-r^2* as the sum-squares here are not mean-corrected. Although the *raw-r^2* satisfies the condition $0 < r^2 < 1$, it is not directly comparable to the conventional r^2 and the interpretations of the two r^2s (for intercept-present and intercept-absent models) are different. For this reason, some scholars do not report the value of r^2 while working with the no-intercept model.

3 The Multiple Linear Regression Model

This chapter extends the discussion of the previous chapter. It is concerned with issues relevant to multiple regression analysis. Specifically, we discuss specification and assumptions of multiple regression model, its estimation, goodness of fit measures, and various problems of inference in the context of multiple regression models. We have added a brief discussion on the LR, Wald, and LM tests which are nowadays widely applied to handle a variety of inference and other problems in multiple regressions. Empirical applications of these tools and techniques have been explained using data set and EViews software package.

3.1 DEFINITION

Sometimes the two-variable regression model may appear to be inadequate as one independent/explanatory variable alone may not adequately explain variation in the dependent variable. In other words, it may appear that there are more than one determinants of the dependent variable. Thus, when we consider more than one determinants or independent variables, it becomes the case of multiple regression models. For instance, if we hypothesise that monthly consumption expenditure of the people is determined by their income, age, education, sex, etc., we have to specify a multiple regression model. In brief, a multiple regression model is the one where two or more independent variables are considered to explain variation in the dependent variable.[1]

Obviously, the easiest example of a multiple regression model is where only two independent variables or regressors are considered. In this chapter, we consider such a model while in the appendix to this chapter we present the multiple regression model involving more than two independent variables.

[1] Geweke et al. (2008, 610) observed that R. Benini, the Italian statistician, was the first to make use of the method of multiple regression in economics in the decade beginning 1900. However, Henry Moore was the first to place the statistical estimation of economic relations at the centre of quantitative analysis of economics in the 1910s. Moore is also credited for laying the foundation of 'statistical economics', the precursor of econometrics.

3.2 SPECIFICATION AND ASSUMPTIONS

The three-variable population regression model involving the dependent variable Y_i and independent/explanatory variables X_{1i} and X_{2i} is specified as:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{3.1}$$

Here ε_i is the stochastic disturbance term and the subscript i denotes the i^{th} observation. As in the case of two-variable model, we make the following assumptions in context of the above multiple regression model.

(i) Zero mean of ε_i: $E(\varepsilon_i \mid X_{1i}, X_{2i}) = 0$ for each i
(ii) Homoskedasticity: $Var(\varepsilon_i) = \sigma^2$ constant
(iii) Non-autocorrelation: $Cov(\varepsilon_i, \varepsilon_j) = 0$ where $i \neq j$
(iv) Normality: ε_i is normally distributed.
(v) Non-stochastic Xs, which implies that the values of the X-variables are same in repeated samples.
(vi) Zero covariance between ε_i and X variables, i.e., $Cov(\varepsilon_i, X_{1i}) = Cov(\varepsilon_i, X_{2i}) = 0$.
(vii) No exact linear relationship exists between the X variables, i.e., Xs are not correlated.

3.3 OLS ESTIMATION

To obtain the OLS estimates of parameters of the population regression model (3.1), let us write the corresponding sample regression model as

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + e_i \tag{3.2}$$

where e_i represents estimated residual values, and $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are estimates of population parameters α, β_1, and β_2, respectively. As in the two-variable model, we apply the 'least-squares criterion' to obtain these estimates. Following this criterion, we select the values of $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ which minimize $\sum e_i^2$.

Here

$$\sum e_i^2 = \sum (Y_i - \hat{\alpha} - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

The necessary conditions of minimization of $\sum e_i^2$ are

$$\frac{\partial \sum e_i^2}{\partial \hat{\alpha}} = \frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = \frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = 0$$

Applying these conditions, the following 'normal equations' are obtained.

$$\sum Y_i = n\hat{\alpha} + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} \tag{3.3a}$$

$$\sum X_{1i}Y_i = \hat{\alpha}\sum X_{1i} + \hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{1i}X_{2i} \tag{3.3b}$$

$$\sum X_{2i}Y_i = \hat{\alpha}\sum X_{2i} + \hat{\beta}_1 \sum X_{1i}X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 \tag{3.3c}$$

It is clear that with the given data on Y_i, X_{1i}, and X_{2i}, we have to compute the following quantities to obtain the values of the estimates.

$$n, \sum Y_i, \sum X_{1i}, \sum X_{2i}, \sum X_{1i}Y_i, \sum X_{2i}Y_i, \sum X_{1i}X_{2i}, \sum X_{1i}^2, \text{ and } \sum X_{2i}^2$$

Putting these values in the aforementioned 'normal equations' and solving, we have solutions for values of $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$. Using these values, we write the estimated three-variable multiple regression model as

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

An alternative way to compute the estimates is to use the following formulas that can be derived by solving the 'normal equations'.

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \tag{3.4}$$

$$\hat{\beta}_1 = \frac{\sum x_{1i}y_i \sum x_{2i}^2 - \sum x_{2i}y_i \sum x_{1i}x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} \tag{3.5}$$

$$\hat{\beta}_2 = \frac{\sum x_{2i}y_i \sum x_{1i}^2 - \sum x_{1i}y_i \sum x_{1i}x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2} \tag{3.6}$$

Here $\bar{Y}, \bar{X}_1$, and $\bar{X}_2$ denote sample mean values for the three variables and the lowercase letters denote deviation from these sample means.

It is also easy to compute the variances of $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ by using the following formulas.

$$Var(\hat{\alpha}) = \left[\frac{1}{n} + \frac{\bar{X}_1^2 \sum x_{2i}^2 + \bar{X}_2^2 \sum x_{1i}^2 - 2\bar{X}_1\bar{X}_2 \sum x_{1i}x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2}\right]\sigma^2 \tag{3.7}$$

$$Var(\hat{\beta}_1) = \left[\frac{\sum x_{2i}^2}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i}x_{2i})^2}\right]\sigma^2 = \frac{\sigma^2}{\sum x_{1i}^2(1 - r_{12}^2)} \tag{3.8}$$

$$Var(\hat{\beta}_2) = \left[\frac{\sum x_{1i}^2}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i}x_{2i})^2}\right]\sigma^2 = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{12}^2)} \tag{3.9}$$

In the above formulae, r_{12} is the sample coefficient of correlation between X_{1i} and X_{2i}, σ^2 is variance of the disturbance term ε_i, which is estimated as

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-3} \tag{3.10}$$

where 3 is the number of parameters in the population regression equation estimated in the model.

3.4 PROPERTIES OF OLS ESTIMATORS

As in two-variable linear model, the least-squares estimators ($\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$) in the three-variable model are also BLUE, i.e., best, linear and unbiased estimators of population parameters (α, β_1, and β_2). It is easy to prove these properties. However, we skip this exercise here as Appendix 3.1 provides proof of BLUE properties in context of the general linear multiple regression model.

3.5 MEASURING GOODNESS OF FIT

After estimating the multiple regression model, we may be interested to assess the goodness or quality of fit of our estimated model. In other words, our objective is to know how well the estimated line fits the sample observations.

The goodness of fit of the estimated model in the context of a two-variable model is understood in terms of the value of r^2-statistic. To recapitulate, r^2-statistic provides a measure of proportion of total variation in the dependent variable that is explained by the independent/explanatory variable of the model. We can extend this concept further to obtain a measure of goodness of fit of the estimated model in the context of estimated multiple regression model. This is done as follows.

In the two-variable model,

$$\begin{aligned} r^2 &= \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \\ &= \frac{\hat{\beta} \sum x_i y_i}{\sum y_i^2} \\ &= \frac{ESS}{TSS} \end{aligned}$$

Let us rewrite the above relation supposing that the variables considered are Y_i and X_{1i}. Then,

$$r^2 = \frac{\hat{\beta}_1 \sum x_{1i} y_i}{\sum y_i^2}$$

Now if we suppose that there are two explanatory variables, X_{1i} and X_{2i}, then

$$R^2 = \frac{\hat{\beta}_1 \sum x_{1i} y_i + \hat{\beta}_2 \sum x_{2i} y_i}{\sum y_i^2}$$

The above formula can be extended further by adding terms in the numerator, when we have more than two explanatory variables. If we have k number of explanatory variables in the model, then the R^2 formula becomes

$$R^2 = \frac{\hat{\beta}_1 \sum x_{1i} y_i + \hat{\beta}_2 \sum x_{2i} y_i + \ldots + \hat{\beta}_k \sum x_{ki} y_i}{\sum y_i^2} \tag{3.11}$$

Usefulness of R^2-statistic

R^2-statistic (in brief, R^2) provides a measure of goodness of fit of the estimated multiple regression model to sample data. It also helps to understand the relevance of explanatory variables in the estimated model. The value of R^2 lies between 0 and 1. When the value of R^2 is close to 0, the explanatory variables have not explained much of the variation in the dependent variable of the model and we have a 'bad fit' estimated equation. In other words, we have not considered the explanatory variables that are relevant to explain variation in the dependent variable. On the other hand, when the value of R^2 is high and close to 1, we have a 'good fit' estimated equation, which explains a large part of variation in the dependent variable and the explanatory variables considered in the model are quite relevant.

Misuse of R^2-statistic

In spite of above-mentioned usefulness of the R^2-statistic, one must be cautious about its possible misuses. In particular, it is to be remembered that it is dangerous to play the game of maximizing the value of R^2. Some researchers do this by gradually increasing the number of explanatory variables in the model. However, in empirical research, quite often we come across a situation where the value of R^2 is high but very few of the estimated coefficients are statistically significant and/or they have expected signs. Therefore, the researchers should be more concerned about the logical/theoretical relevance of the explanatory variables to the dependent variable and also their statistical significance. If in this process, a high value of R^2 is obtained, well and good. On the other hand, if R^2 is low, it does not mean that the model is necessarily bad, particularly when a good number of the estimated coefficients have expected signs and are statistically significant.

To illustrate the above point further, let us consider an interesting example given by Rao and Miller (1972, 14–16) which clarifies the difficulty of choosing between two different models solely on the basis of their computed R^2 values. Rao and Miller estimated both the

4. AUTOCORRELATION

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

$$Y_t = \alpha + \beta X_t + e_t$$
$$\text{Covariance } (e_t, e_s) = 0 \text{ for } t \neq s$$

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).

X	Y	Y[-1]
1	0.397	0.157
2	0.157	-0.083
3	-0.083	-0.243
4	-0.243	-0.323
5	-0.323	-0.243
6	-0.243	-0.083
7	-0.083	0.077
8	0.077	0.347
9	0.347	