BECC-110

INTRODUCTORY ECONOMETRICS

School of Social Sciences
Indira Gandhi National Open University
Maidan Garhi, New Delhi-110068

EXPERT COMMITTEE

Prof. Atul Sarma (retd.) Former Director Indian Statistical Institute, New Delhi	Prof. M S Bhat (retd.) Jamia Millia Islamia New Delhi	Prof. Gopinath Pradhan (retd.) Indira Gandhi National Open University, New Delhi
Dr. Indrani Roy Choudhury CSRD, Jawaharlal Nehru University New Delhi	Dr. S P Sharma Shyamlal College (Evening) University of Delhi	Prof. Narayan Prasad Indira Gandhi National Open University, New Delhi
Sri B S Bagla (retd.) PGDAV College University of Delhi	Dr. Manjula Singh St. Stephens College University of Delhi	Prof. Kaustuva Barik Indira Gandhi National Open University, New Delhi
Dr. Anup Chatterjee (retd.) ARSD College, University of Delhi	Prof. B S Prakash Indira Gandhi National Open University, New Delhi	Saugato Sen Indira Gandhi National Open University, New Delhi

COURSE PREPARATION TEAM

Block/ Unit Title		Unit Writer
Block 1	**Econometric Theory: Fundamentals**	
Unit 1	Introduction to Econometrics	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 2	Overview of Statistical Concepts	
Unit 3	Overview of Hypothesis Testing	
Block 2	**Regression Models: Two Variables Case**	
Unit 4	Simple Linear Regression Model: Estimation	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 5	Simple Linear Regression Model: Inferences	
Unit 6	Extension of Two Variable Regression Models	Prof. Kaustuva Barik, Indira Gandhi National Open University
Block 3	**Multiple Regression Models**	
Unit 7	Multiple Linear Regression Model: Estimation	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 8	Multiple Linear Regression Model: Inferences	
Unit 9	Extension of Regression Models: Dummy Variable Cases	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi and Prof. B S Prakash, Indira Gandhi National Open University
Block 4	**Treatment of Violations of Assumptions**	
Unit 10	Multicollinearity	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 11	Heteroscedasticity	
Unit 12	Autocorrelation	
Block 5	**Econometric Model Specification and Diagnostic Testing**	
Unit 13	Model Selection Criteria	Dr. Sahba Fatima, Independent Researcher, Lucknow
Unit 14	Tests for Specification Errors	

Course Coordinator: Prof. Kaustuva Barik

Editors: Prof. Kaustuva Barik
Prof. B S Prakash (units 5, 7, 11-12) Saugato Sen (units 7-8)

PRINT PRODUCTION

September, 2021

ISBN:

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi -110068 or visit our website: http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by Registrar, MPDD.

Laser Typeset: Mr. Mukesh Yadav **Cover Design:** Mr. Sandeep Maini **Printed at:**

CONTENTS

COURSE INTRODUCTION

Econometrics is an interface between economics, mathematics and statistics. It is mainly concerned with the empirical estimation of economic theory. The present course provides a comprehensive introduction to basic econometric concepts and techniques. The course is divided into five blocks comprising 14 Units.

Block 1 titled, **Econometric Theory: Fundamentals**, comprises three units. Unit 1 is introductory in nature. It defines econometrics and lists the steps we follow in an econometric study. Unit 2 provides an overview of the concepts frequently used in econometrics. In Unit 3 we define the concept and procedure of hypothesis testing.

Block 2 is titled, **Regression Models: Two Variables Case**. It consists of three Units. Unit 4 begins with the estimation procedure of simple regression model by ordinary least squares (OLS) method. It also describes the properties of OLS estimators and goodness of fit of regression models. Unit 5 continues with the simple regression model and describes the procedure of testing of hypothesis. In this context it explains the procedure of forecasting with regression models. Unit 6 extends the simple regression models in terms of log-linear models and changing the measurement units of the variables in a regression model.

Block 3 titled, **Multiple Regression Models,** considers cases where there are more than one explanatory variable. There are three Units in this Block. Unit 7 deals with estimation of multiple regression models. Unit 8 deals with hypothesis testing in the case of multiple regression models. Unit 9 looks into structural stability of regression models and includes dummy variables as explanatory variables in multiple regression models.

Block 4 deals with **Treatment of Violations of Assumptions**. Unit 10 addresses the issue of multicollinearity. It outlines the consequences, detection and remedial measures of multicollinearity. Unit 11 deals with the issue of heteroscedasticity – its consequences, detection and remedial measures. Unit 12 deals with another important problem in multiple regression models, i.e., autocorrelation. It discusses the consequences, detection and remedial measures of autocorrelation.

Block 5 is titled, **Econometric Model Specification and Diagnostic Testing.** There are two Units in this Block. Unit 13 deals with model selection criteria. In this Unit we discuss issues such as the exclusion of relevant variables and inclusion of irrelevant variables. The subject matter of Unit 14 is tests for specification errors. In this context it gives an outline of Akaike Information Criterion (AIC), Schwarz Information Criterion (SIC), and Mallows' Criterion.

UNIT 1 INTRODUCTION TO ECONOMETRICS *

Structure

1.0 OBJECTIVES

After going through this unit, you will be able to

- explain the significance of econometrics in the field of economics;
- distinguish between econometrics, mathematical economics and economic statistics;
- describe the steps to be followed in an econometric study; and
- distinguish between association and causation.

1.1 INTRODUCTION

Econometrics connects the real world to the existing economic theories. Econometrics is based on the development of statistical methods for testing economic relationships and various economic theories. Econometrics helps us in two ways so far as relationship among variables is concerned: (i) explaining the past relationship among the variables, and (ii) forecasting the value of one variable on the basis of other variables.

Econometrics is an interface between economics, mathematics and statistics. It is mainly concerned with the empirical estimation of economic theories. In a broad sense we can say that it is a branch of social science that combines the tools of mathematics and statistical inferences, and these tools are applied to analyse economic phenomena. Econometrics uses regression technique which establishes an association or relationship between various variables. You should note that such relationships do not imply causation. (i.e., cause and effect relationship). The notion of causation has to originate from some theory of economics.

*Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

1.2 MEANING OF ECONOMETRICS

As mentioned earlier, econometrics deals with 'economic measurement'. It can be defined as a stream of social science which uses techniques of mathematics, statistical inference and economic theory applied to analyze any economic phenomenon. It deals with applications of mathematical statistics to economic data. The objective is to provide empirical support to the economic models constructed with the help of mathematical relationship and therefore obtain numerical results. Thus econometrics makes use of economic theory, mathematical economics, and economic statistics.

Econometrics hence becomes a platform for interaction of economic theory, (microeconomics or macroeconomics) using sophisticated mathematical tools in the form of mathematical equations and economic statistics, that is, data. Economic statistics is developed by collection, processing and presentation of data.

The central concern of mathematical economics is to express economic theory in mathematical forms or equations. These equations are finally are expressed in the form of models. You should note that mathematical economics does not evaluate the measurability or empirical verification of theory.

Economic statistics is primarily concerned with collection, processing and presentation of economic data in the form of charts, diagrams and tables. These data could be on microeconomic variables pertaining to households and firms or it could pertain to macroeconomic variables such as GDP, employment, prices, etc. Data for econometric models could be primary data or secondary data. An economic statistician usually limits himself/ herself to tabulation and processing of data.

Econometrics is mainly interested in empirical verification of economic theories. An econometrician would build models and test economic theories. In mathematical economics the relationship is deterministic. For example,

$$Y_i = a + bX_i \qquad \ldots(1.1)$$

In (1.1) above, Y is the explained variable

X is the explanatory variable

a and b are parameters.

The nature of relationship in econometrics, on the other hand, is stochastic. We add a stochastic error variable u_i to equation (1.1). For example,

$$Y_i = a + bX_i + u_i \qquad \ldots(1.2)$$

We will discuss further in Unit 4 on stochastic relationship among variables. In econometrics we generally require special methods due to the unique nature of economic data since such data are not generated under controlled experiments. The aim of econometrics is to bridge the gap between economic theory and actual measurement simply using the technique of statistical inference.

Thus, you should note three prominent features of econometrics. First, econometrics deals with quantitative analysis of economic relationships. Second, it is based on economic theory and logic. Third, it requires appropriate estimation methods to draw inferences. Thus, if the relationship is not expressed in quantitative terms we cannot apply econometric tools. Further, the variables are related according to some theory or logic; otherwise it will be similar to spurious correlation that you studied in statistics.

1.3 ECONOMICS AND ECONOMETIRCS

In economic theory the statements could be qualitative in nature. On the other hand, as discussed above, econometrics is a composition of mathematical economics, economic statistics and mathematical statistics. Let us take an example. The law of demand states that *ceteris paribus* (i.e., other things remaining the same) a rise in price of a commodity is expected to decrease the quantity demanded of that commodity. Therefore, economic theory predicts a negative or inverse relationship between price and quantity demanded of a commodity.

The law of demand does not provide any numerical measure of the strength of relationship between the two variables namely, price and quantity demanded of the commodity. It fails to answer the question that by how much the quantity will go up or down as a result of a certain change in price of commodity.

Econometrics provides empirical content to most economic theories. The real application of economics in the applied world includes forecasting various crucial economic variables such as sales, interest rates, money supply, price elasticity, etc.

The role of an economist is of great significance for an economy when it comes to understand how the variables would behave over a period of time or how these variables are connected to each other. An economist may be required to assess the impact of a proposed price increase on quantity demanded. For example, the impact of increase in price of electricity can be estimated by an econometrician and the electricity board may increase in price accordingly.

Check Your Progress 1

1) Bring out the differences between econometrics, mathematical economics and statistics.

.....

.....

.....

.....

2) Bring out the prominent features of econometrics.

.....

.....

.....

.....

1.4 METHODOLOGY OF ECONOMETRICS

In econometrics we generally come across several types of economic issues. These issues could be from any branch of economics such as microeconomics, macroeconomics, public economics, international trade, etc. These also could be from any of the sectors of the economy such as agriculture, industry and services. The problem at hand could be different. However, there certain common steps to be followed in an econometric study. These steps are as follows:

1. Construction of a statement of theory or hypothesis
2. Specification of mathematical model of the theory
3. Specification of statistical or econometric model
4. Obtaining requisite data
5. Estimation of the parameters of econometric model
6. Testing of hypothesis
7. Forecasting or prediction
8. Interpretation of results

These eight steps need to be elaborated further. Let us consider an example so that we can comprehend the issues. As you know from introductory macroeconomics, consumption expenditure depends upon income of households. Let us see how an econometric study can be carried out on the above relationship.

Step 1:Construction of a Statement of Theory or Hypothesis

The relationship between consumption and income is complex in nature. There are several factors that that influence consumption expenditure of a household

such as size of family, education level, health status of family members, place of stay (rural/urban), etc. In a simple model, however, the Keynesian consumption function establishes the relationship between consumption expenditure and household income. There are two concepts used by Keynes: average propensity to consume (APC), and marginal propensity to consume (MPC). According to Keynes the APC has a tendency to decline as income level increases. We can take the above statement as a hypothesis. Recall that hypothesis is based on certain theory or logic.

Step 2: Specification of Mathematical Model of the Theory

The consumption function takes the following form:

$$C_i = C_0 + cY_i \quad ...(1.3)$$

The variables C and Y represent consumption expenditure and income respectively. Note that C_0 is autonomous consumption, which is the bare minimum needed for survival. Even if income of a household is zero, consumption will be C_0. We note that for APC to decline, the parameters of equation (1.3) should fulfil the following two conditions: $C_0 > 0$ and $0 < c < 1$. These two conditions will help us in formulation of hypothesis in mathematical form.

Step 3: Specification of Statistical or Econometric Model

The consumption income relationship specified in equation (1.3) is exact in nature. If we plot the graph for equation (1.4) we will obtain a straight line. As mentioned earlier, the nature of relationship in econometrics is *stochastic*. Let us consider two households with the same level of income. Their consumption expenditure would be different due to certain factors other than income (such as health status of family members). In order to incorporate such factors we include another variable, u_i, in our model. The variable u_i has to meet certain conditions (to be discussed in Unit 4). Thus the econometric specification of the consumption function would be as follows:

$$C_i = C_0 + cY_i + u_i \quad ...(1.4)$$

Step 4: Obtaining Requisite Data

Data can be obtained from primary sources or secondary sources. You should refer to Unit 1 of the course BECC 107: Statistical methods for Economics for details on primary data and secondary data. In that Unit we have discussed the procedure of conducting sample survey and the important sources of secondary data.

For estimation of our econometric model given at equation (1.4) we need data on two variables, viz., income (Y) and consumption expenditure (C). As you know, income and expenditure are flow variables. Thus we have to specify a time period for these variables. For convenience from measurement point of view, we can take monthly income and monthly expenditure. Second, we have to define

what constitutes a household – who all are members of a household and who all are not included in the household. Third, we have to decide on the nature of data we collect.

As you know, four types of data are available. (i) time series, (ii) cross- sectional, (iii) pooled-data, and (iii) panel data.

(i) Time Series

Time series data are collected on a variable regularly over a period of time. There are some variables on which data is available on a daily basis (e.g., SENSEX and NIFTY). In the case of some other variables, it is available on monthly basis (e.g., consumer price index), on a quarterly basis (e.g., GDP) or on an annual basis (e.g., fiscal deficit).

(ii) Cross-Sectional Data

Cross-sectional data refers to data on several variables at a point of time. For example through a sample survey we can collect household data on expenditure, income, saving, debt, etc. Remember that time series data focuses on the same variable over a period of time while cross-sectional data focuses on several variables at the same point of time. Census data is an example of cross-sectional data.

(iii) Pooled Data

In the pooled data we have elements of both the time series and cross-sectional data. It is a time series of cross-sections. The observations in each cross section may not refer to the same unit. Let us consider an example. The census data in India is collected decennially. The number of households in each census however differs. Such data can be pooled to analyse the shifts in population characteristics over time. You can think of several other examples of pooled data. Examples could be employment and unemployment surveys, workforce participation rates, human development index, etc.

(iv) Panel Data

It is a special type of pooled data. Here observations are taken on the same sample units at multiple points of time. Suppose we want to analyse the variability of returns across shares in the stock market. We can take a sample of 50 public limited companies and observe their net asset value (NAV) daily for the month of August 2021. Thus we get 31 cross sections (since the month August has 31 days) of 50 firms. This constitutes a panel data. We call it a 'balanced panel' if all observations (for time period 1 to t; and for sample units 1 to n) are available. We call it an 'unbalanced panel' if some observations are missing.

Step 5: Estimation of the Parameters of the Econometric Model

We have discussed about sampling procedure, statistical estimation and testing of hypothesis in Block 4 of BECC 107. You need a thorough understanding of those concepts. Remember that in econometric estimation, the number of equations is more than the number of parameters. In order to estimate such models we need certain estimation methods. As you will come to know in subsequent Units of this course, there are quite a few estimation methods. You have been introduced to the least squares method in Unit 5 of the course BECC 107: Statistical Methods for Economics. There are certain econometric software available for estimation purpose. You will learn about econometric software in the course BECE 142: Applied Econometrics.

Step 6: Testing of Hypothesis

Once you obtain the estimates of the parameters, there is a need for test of the hypothesis. As you know, in a sampling distribution of an estimator, the estimate varies across sample. The estimate that you have obtained could be a matter of chance, and the parameter may be quite different from the estimate obtained. We need to confirm whether the difference between the parameter and the estimate really exists or it is a matter of sampling fluctuation.

For the consumption function (1.4), we should apply one sided t-test for testing of the condition $C_0 > 0$. For the marginal propensity to consume we should apply two-sided t-test $H_0 : c = 0$. For testing both the parameters together we should apply F-test.

There is a need to check for the correct specification of the model. Two issues are important here: (i) how many explanatory variables should be there in the regression model, and (ii) what is the functional form of the model.

The consumption function (see equation (1.4)) is a case of two-variable regress model. There is one explained variable and one explanatory variable in the model. If we include more number of explanatory variables (such as education, type of residential area, etc.) it becomes a multiple linear regression model. The functional form again could be linear or non-linear.

Step 7: Forecasting or Prediction

The estimated model can be used for forecasting or prediction. We have the actual value of the dependent variable. On the basis of the estimated regression model, we obtain the predicted value of the dependent variable. The discrepancy between the two is the prediction error. This prediction error is required to be as small as possible.

Step 8: Interpretation of Results

There is a need for correct interpretation of the estimates. In later Units of this course we will discuss issues such as model specification and interpretation of the result. The estimated model can be used for policy recommendation also.

1.5 ASSOCIATION AND CAUSATION

As you know from 'BECC 107: Statistical Methods for Economics' correlation implies association between two variables. Technically we can find out the correlation coefficient between any two variables (say the number of students visiting IGNOU library and the number of road accidents in Delhi). In some cases we find the correlation coefficients to be high also. Such relationship between variables however leads to spurious correlation. If we take two such variables (where correlation coefficient is high) and carry out a regression analysis we will find the estimates to be statistically significant. Such regression lines are meaningless. Thus regression analysis deals with the association or dependence of one variable on the other. It does not imply 'causation' however. The notion of causation has to come from existing theories in economics. Therefore a statistical relationship can only be statistically strong or suggestive. Unless causality is established between the variables the purpose of testing the economic theory would not make any sense. Most of the economic theories test the hypothesis whether one variable has a causal effect on the other.

Thus logic or economic theory is very important in regression analysis. We should not run a regression without establishing the logic for the relationship between the variables. Let us look into the case of the law of demand. While analysing consumer demand, we need to understand the effect of changing price of the good on the quantity demanded holding the other factors such as income, price of other goods, tastes and preferences of individuals unchanged. However, if the other factors are not held fixed, then it would be impossible to know the causal effect of price change on quantity demanded.

Check Your Progress 2

1) Explain the steps you would follow in an econometric study.

.....

.....

.....

.....

.....

2) Assume that you have to carry out an econometric study on Keynesian consumption function. Write down the steps you would follow.

.....

.....

.....

.....

.....

3) What do you understand by cause and effect relationship? How is it different from association?

.....

.....

.....

.....

.....

1.6 LET US SUM UP

In this Unit we dealt with the significance of econometrics in the field of economics. Econometrics connects the real world with theory. It helps us to ascertain the validity of theory.

Behind every econometric model there should be certain logic. The relationship between variables should come from certain economic theory or logic. Mere estimation of a regression model may give up meaningless results.

In this Unit we described the steps of carrying out econometric analysis. There are eight steps that we should follow while conducting an econometric study.

1.7 ANSERS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) In Section 1.2 we have shown that econometrics and interface between economics, statistics and mathematical economics. Elaborate on that.

2) There are three prominent features of econometrics. First, econometrics deals with quantitative analysis of economic relationships. Second, it is based on economic theory and logic. Third, it requires appropriate estimation methods to draw inferences.

Check Your Progress 2

1) You should explain the eight steps mentioned in Section 1.4.

2) You should follow the eight steps given in Section 1.4. Your answer may include the following:

   (i) Statement of the theory: $0< MPC<1$

   (ii) Mathematical specification of the model: $C = \beta_1 + \beta_2 Y$, $0< \beta_2 <1$

   (iii) Econometric specification the model: $C = \beta_1 + \beta_2 Y + u$

   (iv) Collection of Data: Secondary data from RBI Handbook of Statistics

   (v) Parameter Estimation: $\widehat{C_i} = -184.08 + 0.7164Y_i$

   (vi) Hypothesis Test: $\beta_1 > 0$ or $\beta_2 > 0$

   (vii) Prediction: what is the value of C, given the value of Y?

3) Regression analysis deals with the association or dependence of one variable on the other. It does not imply causation. The notion of causation has to come from outside statistics. It could be some existing theory in economics. Therefore a statistical relationship can only be statistically strong or suggestive. Most of the economic theories test the hypothesis whether one variable has a causal effect on the other. Regression *per se* is all about association between two or more variables; this association might be suggestive. Unless causality is established between the variables the purpose of testing the economic theory would not make any sense.

UNIT 2 OVERVIEW OF STATISTICAL CONCEPTS*

Structure

2.0 OBJECTIVES

After going through this unit, you will be able to

- explain the concept and significance of probability distribution;
- identify various types of probability distributions;
- describe the properties of various probability distributions such as normal, t, F and chi-square;
- explain the process of estimation of parameters and
- describe the properties of a good estimator.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

2.1 INTRODUCTION

Statistical concepts and estimation methods hold crucial significance in understanding the tools of econometrics. Therefore, you should be able to define various concepts and distinguish between them. The essence of econometrics is based on empirical analysis which deals with data. In fact, the tools of econometric analysis emerge from statistical methods.

Statistical concepts guide us to make judgement in the presence of uncertainty. Statistics provides the platform for data collection methods which becomes the basis for carrying out econometric analysis. Econometricians need to work with large population, which becomes a challenge. Therefore, there is a need to select appropriate sample and draw appropriate inferences based on probability distributions. Econometrics calls for a strong understanding of statistical concepts which help economists to choose the right sample and infer correctly from the chosen sample.

The population is a collection of items, events or people. It is difficult to examine every element in the population. Therefore it makes sense in taking a subset of the population and examining it. This subset of population is called a 'sample' which is further used to draw inferences. If the sample is random and large enough, the information collected from the sample can be used for making inference about the population.

Any experiment which gives random outcomes is referred to as random experiment. A variable which takes values which are outcome of random process is called a random variable. Thus, for a random variable each outcome is associated with certain probability of occurrence.

Random variables are discrete random variables when they take finite values. If the random variable assumes infinite number of values between any two pints, it is called a continuous random variable. Random variables have a probability distribution. If the random variable is discrete then the probability function associated with it is called 'probability distribution function'. If the random variable is continuous, then the probability function is referred to as 'probability density function'. Random variables can have variety of distribution functions depending on their probabilities. Some of the commonly used distribution functions are described in this Unit.

2.2 STATISTICAL INFERENCE

In BECC 107 we have discussed the procedure of statistical inference in detail (You should go through Units 13 and 14 of BECC 107). Statistical inference is the method of drawing conclusions about the population characteristics on the basis of information contained in a sample drawn from the population. Remember that population mean is not known to us, but we know the sample

mean. In statistical inference we are interested in answering two types of questions. First, what would be the value of the population mean? The answer lies in making an informed guess about the population mean. This aspect of statistical inference is called 'estimation'. The second question pertains to certain assertion made about the population mean. Suppose a manufacturer of electric bulbs claims that the mean life of electric bulbs is equal to 2000 hours. On the basis of the sample information, can we say that the assertion is not correct? This aspect of statistical inference is called hypothesis testing. Thus statistical inference deals with two issues: (i) estimation, and (ii) hypothesis testing. We discuss about estimation of parameters in the present Unit. Hypothesis testing will be discussed in Unit 3.

If expected Price-Earning Ratio of 28 companies is 23.25, then this sample average can be used as an estimate of the population average of stocks. As you know, the sample average (or, sample mean) is denoted by $\bar{X}$. This sample mean can be inferred as the expected value of X, which is the population mean. This process of generalizing from the sample value ($\bar{X}$) to the population value E(X) is the essence of statistical inference.

Statistical inference aims at understanding the characteristics of population from the sample. These population characteristics are the 'parameters' of the population and the characteristics of the sample are the 'statistics'. The method of determining and computing population parameter using the sample is called *estimation*.

2.3 CENTRAL LIMIT THEOREM

When the functions of random variables are independent and identically distributed then as the sample size increases, the sample mean tends to be normally distributed around the population mean and the standard deviation reduces as sample size 'n' increases.

If X_1, X_2, X_3, ……. and X_n are independent and identically distributed with mean μ and standard deviation σ, then sample mean ($\bar{X}$) is given by

$$\bar{X} = \frac{(X_1 + X_2 + \ldots\ldots + X_n)}{n} \qquad \ldots(2.1)$$

The central limit theorem implies that the expected sample mean and standard deviation (SD) would converge as follows:

$$E(\bar{X}) = \mu \text{ and } SD(\bar{X}) = \frac{\sigma}{\sqrt{n}} \qquad \ldots (2.2)$$

The Central Limit Theorem (CLT) states that

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N(0, 1) \text{ as } n \rightarrow \infty \qquad \ldots (2.3)$$

This further implies that $\bar{X} \rightarrow N(\mu, \frac{\sigma^2}{n})$. In other words, the sample mean can be approximated with a normal random variable with mean μ and standard deviation $\frac{\sigma^2}{n}$. We discuss certain important probability distribution functions below.

2.4 NORMAL DISTRIBUTION

Normal distribution (also called z-distribution) is a continuous probability distribution function. This function is very useful because of Central Limit Theorem. It implies that averages of samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal. It becomes normally distributed when the number of observations is sufficiently large. The normal distribution is also called the bell curve (see Fig. 2.1). The probability density function (pdf) of normal distribution is

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \ldots (2.4)$$

where, μ is the expectation of distribution or mean

σ is the standard deviation, and σ^2 is the variance.

Some of the important properties of normal distribution are:

a) The normal distribution curve is bell-shaped.

b) The normal curve is symmetrical about the mean μ.

c) The total area under the curve is equal to 1.

d) The area of the curve is completely described by its mean and standard deviation.

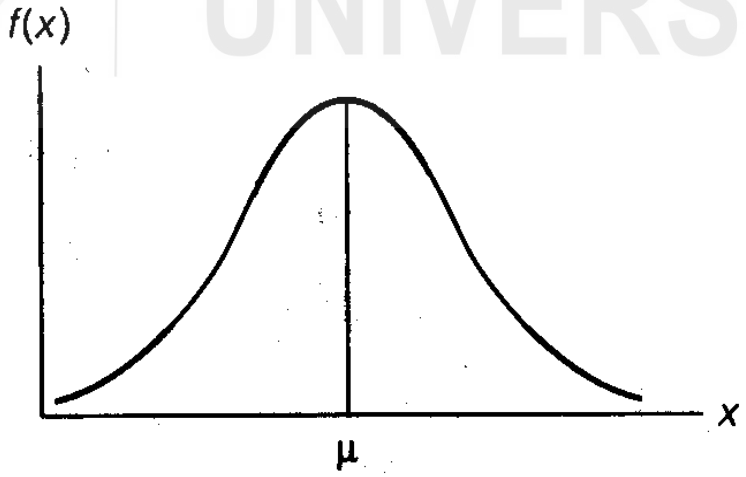


Fig. 2.1: Normal Probability Distribution

Standard Normal Distribution: N ~ N (0,1)

It is a normal distribution with mean zero ($\mu = 0$) and unit variance ($\sigma^2 = 1$), then the probability distribution function is given by

$$f(x \mid 0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{... (2.5)}$$

All the properties of the normal distribution mentioned above are applicable in the case of standard normal distribution.

Check Your Progress 1

1) Assume that X is normally distributed with mean μ = 30 and standard deviation σ = 4. Find P(X < 40).

.....

.....

.....

.....

.....

.....

2) Bring out the important properties of normal probability distribution.

.....

.....

.....

.....

.....

.....

3) The life of an electronic bulb produced by a company follows normal distribution with mean of 12 months and standard deviation of 2 months. Find out the probability that a bulb produced by the company will last

a) less than 7 months

b) between 7 and 12 months

.....

.....

.....

.....

.....

.....

2.5 CHI-SQUARE DISTRIBUTION

Suppose X is a normal variable with mean μ and standard deviation σ, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variable, i.e., $z \sim N(0,1)$. If we take the square of z, i.e., $z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$, then z^2 is said to be distributed as a χ^2 variable with one degree of freedom and expressed as χ_1^2.

It is clear that since χ_1^2 is a squared term; for z laying between $-\infty$ and $+\infty$, χ_1^2 will lie between 0 and $+\infty$ (because a squared term cannot take negative values). Again, since, z has a mean equal to zero, most of the values taken by z will be close to zero. As a result, the probability density of χ_1^2 variable will be maximum near zero.

Generalizing the result mentioned above, if $z_1, z_2, \ldots, z_k$ are independent standard normal variables, then the variable

$$z = \sum_{i=1}^{ki} z_1^2$$

is said to be a χ^2 variable with k degrees of freedom and is denoted by χ_k^2. Fig. 2.2 given below shows the probability curves for χ^2 variables with different degree of freedom.

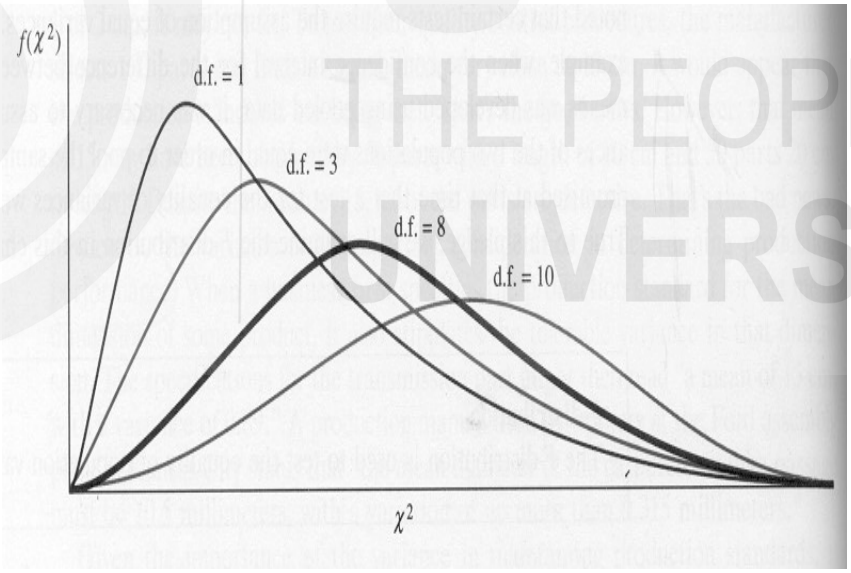


Fig. 2.2: Chi-Square Probability Curves

The chi-square distribution is one of the most widely used probability distributions. The area under the chi-square probability curve is equal to 1.

Unlike standard normal distribution, the distribution of chi-square changes its shape with sample size. In the case of small samples, the distribution is skewed to

the right but it becomes symmetric as the sample size increases. All the values of the chi-square distribution are positive.

Properties of Chi-square distribution

1. The mean of the chi-square distribution is equal to the number of degrees of freedom (k).
2. The variance of the chi-square distribution is equal to two times the number of degrees of freedom: $\sigma^2 = 2k$
3. When the degrees of freedom are greater than or equal to 2, the maximum value of Y occurs when $\chi^2 = k - 2$.
4. As the degree of freedom increases, the chi-square curve approaches a normal distribution.

2.5 THE *t*- DISTRIBUTION

The t-distribution is also called the student's t-distribution. It was introduced by the English statistician W S Gosset under the penname 'Student'. It belongs to the family of continuous probability distributions. The t-distribution is applicable the sample size is small and population standard deviation is unknown. The cases where population parameters, i.e., μ and σ are not known and are estimated using sample statistics. The t-distribution is symmetric as in the case of the standard normal distribution (z). The height of the t-distribution depends on the sample size (see Fig. 2.3). As n approaches ∞, the t-distribution approaches the standard normal distribution.

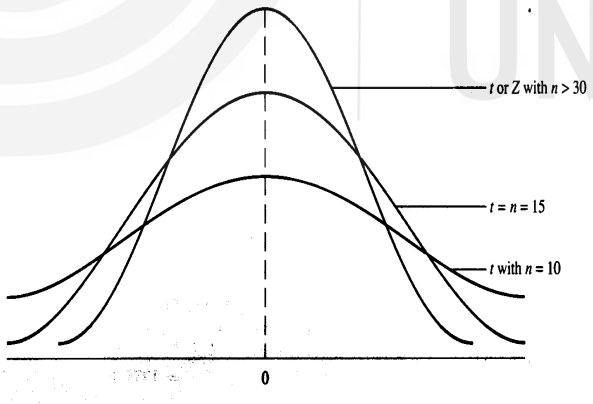


Fig 2.3: Student's-t Probability Curves

If z_1 is a standard normal variate, i.e., $z_1 \sim N(0,1)$ and z_2 is another independent variable that follows the chi-square distribution with k degrees of freedom, i.e., $z_2 \sim \chi^2_k$, then the variable

$$t = \frac{z_1}{\sqrt{(z_2/k)}} = \frac{z_1\sqrt{k}}{\sqrt{z_2}} \quad \ldots (2.6)$$

is said to follow student's-*t* distribution with *k* degrees of freedom.

The value of t-distribution can be obtained as:

$$t = \frac{[\bar{X} - \mu]}{[\frac{s}{\sqrt{n}}]} \quad \ldots (2.7)$$

where, $\bar{X}$ is the sample mean, μ is the population mean, s is the standard deviation of the sample and *n* is the sample size.

Properties of *t*-Distribution

1. The mean of the distribution is equal to 0
2. The variance is equal to [k/(k – 2)] where k is the degrees of freedom and k $\geq$2.
3. The variance is always greater than 1, although it is close to 1 when the degree of freedom is large. For infinite degrees of freedom the t-distribution is the same as the standard normal distribution.

The t-distribution can be used under the following conditions:

1. The population distribution is normal
2. The population distribution is symmetric, unimodal without outliers, and the sample size is at least 30
3. The population distribution is moderately skewed, unimodal without outliers and the sample size is at least 40
4. The sample size is greater than 40 without outliers.

Look into the above conditions. If the parent population (from which the sample is drawn) is normal we can apply t-distribution for any sample size. If population is not normal, the sample size should be large. The t-distribution should not be used with small samples drawn from a population that is not approximately normal.

2.7 THE *F*- DISTRIBUTION

Another continuous probability distribution that we discuss now is the *F* distribution. If z_1 and z_2 are two chi-squared variables that are independently distributed with $k_1$and k_2 degrees of freedom respectively, the variable

$$F = \frac{z_1/k_1}{z_2/k_2} \quad \ldots (2.8)$$

follows *F* distribution with k_1 and k_2 degrees of freedom respectively. The variable is denoted by F_{k_1,k_2} where, the subscripts k_1 and k_2 are the degrees of freedom associated with the chi-squared variables.

You should note that k_1 is called the numerator degrees of the freedom and in the same way, k_2 is called the denominator degrees of freedom.

Some important properties of the F distribution are mentioned below.

1) The *F* distribution, like the chi-squared distribution, is also skewed to the right. But, as k_1 and k_2 increase, the *F* distribution approaches the normal distribution.

2) The mean of the *F* distribution is $k_1/(k_2 - 2)$, which is defined for $k_2 > 2$, and its variance is $\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$ which is defined for $k_2 > 4$.

3) An *F* distribution with 1 and *k* as the numerator and denominator degrees of freedom respectively is the square of a student's-*t* distribution with *k* degrees of freedom. Symbolically,

$$F_{1,k} = t_k^2$$

4) For fairly large denominator degrees of freedom k_2, the product of the numerator degrees of freedom k_1 and the *F* value is approximately equal to the chi-squared value with degrees freedom k_1, i.e., $k_1 F = \chi^2_{k_1}$.

The *F* distribution is extensively used in statistical inference and testing of hypotheses. Again, such uses also require obtaining areas under the *F* probability curve and consequently integrating the *F* density function. However, in this case also our task is facilitated by the provision of the *F* Table.

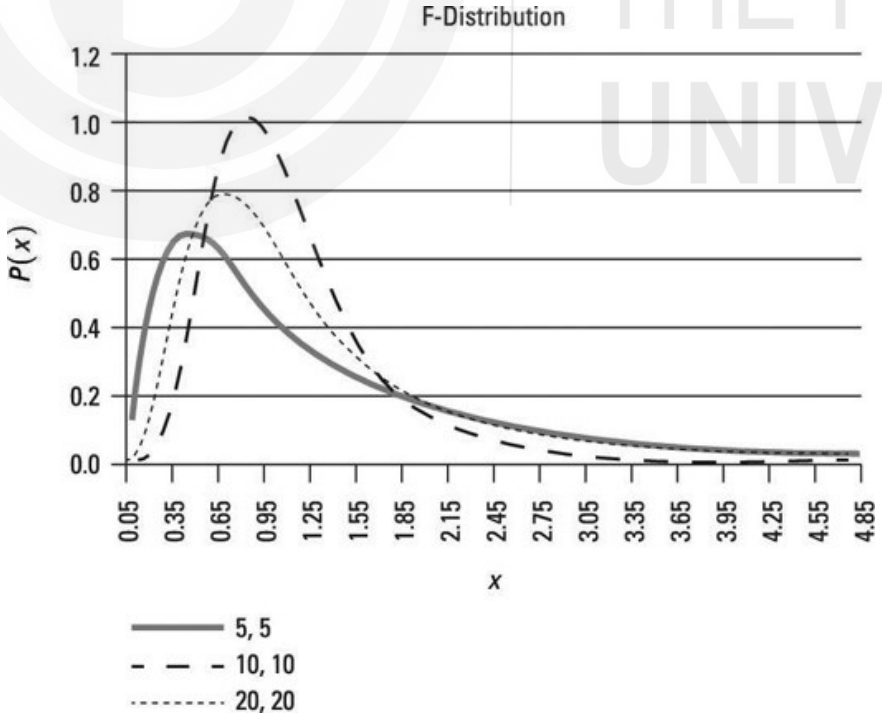


Fig. 2.4: Probability Curves of *F*-Distribution

The F-distribution is used to test the population variance. We can test whether two normal populations have the same variance. The null hypothesis is that the variances are same while alternative hypothesis is that one of the variances is larger than the other. That is:

$$H_{o:}\ \sigma_1^2 = \sigma_2^2$$

$$H_{A:}\ \sigma_1^2 > \sigma_2^2$$

The alternative hypothesis states that the first population has larger variance. The null hypothesis can be tested by drawing a sample from each population and calculating the estimates s_1^2 and s_2^2. The samples are assumed to be independently drawn with size n_1 and n_2 respectively. We test the ratio

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_{1-1,,}n_2-1} \qquad \ldots (2.9)$$

If null hypothesis is not true, the ratio would be statistically different from unity. We should compare the calculated value of F (obtained from equation (2.9)) with the tabulated value of F (given in the appendix table at the end of the book). If the calculated value exceeds the tabulated value, then the null hypothesis is rejected.

Check Your Progress 2

1) A newly developed battery lasts 60 minutes on single charge. The standard deviation is 4 minutes. For the purpose of quality control test, the department randomly selects 7 batteries. The standard deviation of selected batteries is 6 minutes. What is the probability that the standard deviation in new test would be greater than 6 minutes?

.....

.....

.....

.....

.....

.....

2) Define chi-square distribution. Bring out its important properties.

.....

.....

.....

.....

.....

.....

3) Suppose the scores on a GRE test are normally distributed with population mean of 100. Suppose 20 people are randomly selected and tested. Sample standard deviation is 15. What is the probability that the average test score will be at most 110?

.....

.....

.....

.....

.....

.....

4) Test whether the students from private high schools are more homogeneous with respect to their science test score than the students from public high schools. It is given that the sample variances are 91.74 and 67.16 respectively for public and private schools. The sample sizes of the students are 506 for public schools and 94 for private schools.

.....

.....

.....

.....

.....

2.8 ESTIMATION OF PARAMETERS

Estimation could be of two types: (i) point estimation, and (ii) interval estimation. In point estimation we estimate the value of the population parameter as a single point. On the other hand, in the case of interval estimation, we estimate the lower and upper bounds around the sample mean within which the population mean is likely to remain.

2.8.1 Point Estimation

Let us assume that a random variable X follows normal distribution. As you know, normal distribution is described by two parameters, viz., mean and standard deviation. Since we do not have data for the whole population (we have data for a sample only), we need to estimate mean E(X) = μ_X and variance σ_X^2 on the basis of a sample only.

Let us assume that we have data from a random sample of size *n* (suppose, sample size n = 50) from a known probability distribution (say, normal distribution). We use the sample to estimate the unknown parameters. Suppose, we find sample mean $\bar{X}$ to be 23.28. This single numerical value is called the

point estimate of the parameter where $\bar{X} = \frac{\Sigma X_i}{n}$. This formula is called the point estimator. You should note that the point estimator is a random variable as its value varies from sample to sample.

2.8.2 Interval Estimation

In point estimation we estimate the parameter by a single value, usually the corresponding sample statistic. The point estimate may not be realistic in the sense that the parameter value may not exactly be equal to it.

An alternative procedure is to give an interval, which would hold the parameter with certain probability. Here we specify a lower limit and an upper limit within which the parameter value is likely to remain. Also we specify the probability of the parameter remaining in the interval. We call the *interval* as 'confidence interval' and the *probability* of the parameter remaining within this interval as 'confidence level' or 'confidence coefficient'.

The concept of confidence interval is somewhat complex. We have already explained it in BECC 107, Unit 13. Let us look at it again. We have drawn a sample of size *n* from a normal population. We do not know the population mean μ_X and population variance σ_X^2. We know the sample mean $\bar{X}$ and sample variance S_X^2. Since $\bar{X}$ varies across samples, we use the properties of the sampling distribution of $\bar{X}$ to draw inferences about μ_X.

If X is normally distributed, i.e. , we know that

$$\bar{X} \sim \left(\mu_X, \frac{\sigma_x^2}{n} \right) \quad \ldots (2.10)$$

From (2.10) we can say that sampling distribution of sample mean $\bar{X}$ follows normal distribution with mean μ_X and standard deviation σ_X^2/n. Let us transform the above as a standard normal variable.

$$Z = \frac{\bar{X} - \mu_x}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0, 1) \quad \ldots (2.11)$$

Now the problem before us is that we do not know the population variance σ_X^2. Thus we take its estimator $S_X^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$. In that case, the appropriate test statistic is

$$t = \frac{(\bar{X} - \mu_X)}{S_x / \sqrt{n}} \quad \ldots (2.11)$$

Equation (2.11) follows *t*-distribution with $(n-1)$ degrees of freedom.

By re-arranging terms in equation (2.11) we obtain the confidence interval of μ_X.

This also helps us to obtain an interval estimation of μ_X.

For 27 degrees of freedom (d.f.), the tabulated value is 2.052 at the 5 per cent level of significance (see Appendix Table). Thus

$P(-2.052 \le t \le 2.052) = 0.95$ … (2.12)

The critical t values show the percentage of area under the t-distribution curve that remains between those values. The value $t = -2.052$ is called the lower critical value, and the value $t = 2.052$ is called the upper critical value.

Equation (2.12) implies that for 27 d.f. the probability is 0.95 or 95% that the interval (–2.052, 2.052) will include μ_X.

$$\therefore P\left(-2.052 \le t = \frac{\bar{X} - \mu_X}{S_x/\sqrt{n}} \le 2.052\right)$$

$$\Rightarrow \quad P\left(\bar{X} - 2.052\frac{S_X}{\sqrt{n}} \le \mu_X \le \bar{X} + 2.052\frac{S_X}{\sqrt{n}}\right) = 0.95 \qquad (2.13)$$

Equation (2.13) provides an interval estimator of μ_X. It is called the 95% confidence interval (CI) for the true but unknown population mean μ_X. The value 0.95 is called the confidence coefficient. It implies that the probability is 0.95 that random interval $\bar{X} \pm 2.052\frac{S_X}{\sqrt{n}}$ contains true μ_X.

$\bar{X} - 2.052\frac{S_X}{\sqrt{n}}$ is called lower limit of interval.

$\bar{X} + 2.052\frac{S_X}{\sqrt{n}}$ is called upper limit of interval.

This is a random interval because the values are based on $\bar{X}$ and $\frac{S_X}{\sqrt{n}}$ which will vary from sample to sample. You should note that μ_X is not random; rather it is a fixed number. Therefore we can say that "the probability is 0.95 that μ_X lies in this interval".

2.9 PROPERTIES OF ESTIMATORS

An estimator is considered as best linear unbiased estimator (BLUE) if it is linear, unbiased, efficient (with minimum variance). and also consistent implying that the the value of estimator converges to its true population value as the sample size increases. All the properties of good estimators are discussed below.

2.9.1 Linearity

An estimator is said to be a linear estimator if it is a linear function of the sample observation

$$\bar{X} = \sum_{i=1}^{n} \frac{Xi}{n}$$

$$= \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right) \quad \ldots (2.14)$$

Sample mean is the linear estimator because it is a linear function of the observations.

2.9.2 Unbiasedness

The value of a statistic varies across samples due to sampling fluctuation. Although the individual values of a statistic may be different from the unknown population parameter, on an average, the value of a statistic should be equal to the population parameter. In other words, the sampling distribution of $\bar{X}$ should have a central tendency towards μ_X. This is known as the property of unbiasedness of an estimator. It means that although an individual value of a given estimator may be higher or lower than the unknown value of the population parameter, there is no bias on the part of the estimator to have values that are always greater or smaller than the unknown population parameter. If we accept that mean (here, expectation) is a proper measure for central tendency, then $\bar{X}$ is an *unbiased estimator* for μ_X if

$$E(\bar{X}) = \mu_X$$

2.9.3 Minimum Variance

An estimator of μ_X is said to be the minimum variance estimator if its variance is smaller than the variance of any other estimator of μ_X. Suppose there are three estimators of μ_X. The variance of $\hat{\mu}_3$ is the smallest of the three estimators. Hence, it is minimum variance estimator.

2.9.4 Efficiency

The property of unbiasedness is not adequate by itself. It is possible to obtain two or more estimators of a parameter as unbiased. Therefore, we must choose the most efficient estimator. Suppose two estimators of μ_X as given as follows:

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right) \quad \ldots (2.15)$$

$$X_{med} \sim N\left(\mu_X, \left(\frac{\pi}{2}\right)\frac{\sigma^2}{n}\right), \qquad \pi = 3.142 \text{ (approx.)} \quad \ldots (2.16)$$

In the case of large samples, the median computed from a random sample of normal population also follows normal distribution with the same μ_X. However, it has a large variance.

$$\frac{Var(\bar{X}_{med})}{Var(\bar{X})} = \frac{\pi}{2}\frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}} - \frac{\pi}{2} = 1.571 \quad \text{(approx.)} \qquad \ldots (2.17)$$

Equation (2.17) implies that the variance of sample median is 57% larger than the variance of sample mean. Therefore, the sample mean provides more precise estimate of population mean compared to the median (X_{med}). Thus, $\bar{X}$ is an efficient estimator of μ_X.

2.9.5 Best Linear Unbiased Estimator (BLUE)

Suppose we consider a class of estimators. Among these estimators, an estimator fulfils three properties, viz., (i) it is linear, (ii) it is unbiased, and (iii) it has minimum variance. In that case, it is called a 'best linear unbiased estimator' (BLUE).

2.9.6 Consistency

Consistency is a large sample property. If we increase the sample size, the estimator should have a tendency to approach the value of the parameter. Thus, an estimator is said to be consistent if the estimator converges to the parameter as $n \to \infty$.

Suppose $X \sim N(\mu_X, \sigma_X^2)$. We draw a random sample of size n from the population.

Two estimators of μ_X are]

$$\bar{X} = \Sigma\frac{X_i}{n} \qquad \ldots (2.18)$$

$$X^* = \Sigma\frac{X_i}{n+1} \qquad \ldots (2.19)$$

As you know, the first estimator (2.18) is the sample mean and it is unbiased since $E(\bar{X}) = \mu_X$.

The second estimator (2.19) is biased as

$$E(X^*) = \left(\frac{n}{n+1}\right)\mu_X$$

Thus, $E(X^*) \neq \mu_X$

As the sample size increases we should not find much difference between the two estimators. As n increases, X^* will approach μ_X. Such an estimator is known as consistent estimator. An estimator is consistent estimator if it approaches the true value of parameter as sample size gets larger and larger.

Check Your Progress 3

1) Describe the desirable properties of an estimator.

.....

.....

.....

.....

2) For a sample of size 30, the sample mean and standard deviation are 15 and 10 respectively. Construct the confidence interval of population mean ($\boldsymbol{\mu_X}$) at 5 per cent level of significance.

.....

.....

.....

.....

2.10 LET US SUM UP

Statistical concepts guide us the way to make judgement in the presence of uncertainty. In this Unit we discussed about certain basic statistical concepts. We discussed about certain continuous probability distributions such as normal, standard normal, chi-square, *t* and *F*. We depicted the probability distribution curves of these curves. In the appendix given at the end of this book, we have given the following: Normal Area Table, and critical values of *t*, chi-square and *F* distributions.

In addition to the above we have described the properties of a good estimator. We have explained concepts such as unbiasedness, consistency and efficiency in the context of an estimator.

2.11 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) You have to find out the area under the standard normal curve.
If X = 40, $z = \frac{(40-30)}{4} = 2.5$
Hence P (X < 40) = P (z < 2.5) = [area to the left of 2.5] = 0.9938

2) Go through Section 2.4 and answer.

3) a) P(X < 7) = P (Z < – 2.5) = 0.0062

b) P(20 < X < 22) = P (–2.5 < Z < 0) = 0.4938

Check Your Progress 2

1) Standard deviation of the population is 4 minutes. Standard deviation of the sample is 6 minutes. The number of sample observations is 7.

   $X^2 = [(n–1)*s^2]/ \sigma^2$

   $X^2 = [(7 – 1) * 6^2] / 4^2 = 13.5$

   Degree of freedom is (n–1) = (7–1) = 6. The probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This implies that the probability that the standard deviation would be greater than 6 minutes is (1 – 0.96) = 0.04.

2) Go through Section 2.5 and answer.

3) Population mean $\mu = 100$. Sample size n = 20. Degrees of freedom is (20–1) = 19. Sample mean $\bar{X}$ should be at most 110. Sample standard deviation s =15. Since we do not know the population standard deviation we apply t-distribution. Applying the formula,

   $t = \frac{[\bar{X} - \mu]}{[\frac{s}{\sqrt{n}}]}$. Thus, $t = \frac{110 - 10}{\frac{15}{\sqrt{20}}} = 0.996$

   This implies 99.6% chance that the sample average will be no greater than 110.

4) The degrees of freedom (n_1–1) and (n_2–1) are 505 and 93 respectively. Our null hypothesis H_0 is that the both type schools are equally homogeneous with respect to science marks. We are comparing variances. Thus we apply F-test.

   $F = {s_1^2}/{s_2^2} = \frac{91.74}{67.16} = 1.366$

   The tabulated value of F for 505 and 93 degrees of freedom is 1.27. Since calculated value is more than the tabulated value, we reject the H_0. We conclude that the students from private schools are more homogeneous with respect to science marks.

Check Your Progress 3

1) Go through Section 2.9 and answer.

2) Since population standard deviation is not known, you should apply t-distribution. Check the tabulated value of *t* given at the Appendix for 29 degrees of freedom and 5 per cent level of significance. Construct the confidence interval as given at equation (2.12).

UNIT 3 OVERVIEW OF HYPOTHESIS TESTING*

Structure

3.0 OBJECTIVES

After going through this unit, you will be able to

- explain the concept and significance of hypothesis testing;
- describe the applications of a test statistic;
- explain the procedure of testing of hypothesis of population parameters;
- distinguish between the Type I and Type II errors; and
- apply the tests for comparing parameters from two different samples.

3.1 INTRODUCTION

The purpose behind statistical inference is to use the sample to make judgement about the population parameters. The concept of hypothesis testing is crucial for predicting the value of population parameters using the sample. Various test statistics are used to test hypotheses related to population mean and variance. The variance of two different samples can also be compared using hypothesis testing. There are two approaches to testing of hypothesis: (i) test of significance

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

approach, and (ii) confidence interval approach. While testing a hypothesis, there is a likelihood of committing two types of errors: (i) type I error, and (ii) type II error. In this unit we will elaborate on the process of hypothesis testing, and explain the method of rejecting the null hypothesis on the basis of appropriate test statistic.

3.2 PROCEDURE OF HYPOTHESIS TESTING

We formulate a hypothesis on the basis of economic theory or logic. A hypothesis is a tentative statement about certain characteristic of a population. As you know, a population is described by its parameters (such as mean, standard deviation, etc.). Thus a hypothesis is an assumption about a population parameter. A hypothesis may or may not be true. For finding out that we test a hypothesis by certain econometric method.

Formulation of a hypothesis involves a prior judgement or expectation about what value a particular parameter may assume. For example, prior knowledge or an expert opinion tells us that the true average price to earnings (P/E) ratio in the local stock exchange is 20. Thus our hypothesis is that the P/E ratio is equal to 20.

In order to test this hypothesis, suppose we collect a random sample of stocks and find that the average P/E ratio is 23. Is the figure 23 statistically different from 20? Because of sampling variation there is likely to be a difference between a sample estimate and its population value. It is possible that statistically the number 23 may not be very different from the number 20. If this is the case, then we should not reject the hypothesis that the average P/E ratio is 20.

In hypothesis testing there are four important components: i) null hypothesis, ii) alternative hypothesis, iii) test statistic, and iv) interpretation of results. We elaborate on these components below.

(i) Formulation of null and alternative hypotheses: There are two types of hypothesis, viz., null hypothesis and alternative hypothesis. A ‘null hypothesis’ is the statement that we consider to be true about the population. It is called ‘null’ thereby meaning empty or void. For example, a null hypothesis could be: there is no relationship between employment and education. Therefore, if we carry out a regression of employment on education, the regression coefficient should be zero. Usually we denote null hypothesis by H_0. The alternative hypothesis is the opposite of the null hypothesis. Alternative hypothesis is usually denoted by H_1. You should note that H_0 and H_1 are ‘mutually exclusive’; they cannot occur simultaneously.

(i) Identification of the test statistic: The null hypothesis is put to test by a test statistic. There are several test statistics (such as t, F, chi-square, etc.) available in econometrics. We have to identify the appropriate test statistic.

(ii) Interpretation of the results based on the value of the test statistic: After carrying out the test, we interpret the results. When we apply the test statistic to the sample data that we have, we obtain certain value of the test statistic (for example, t-ratio of 2.535). Interpretation of results involves comparison of two values: tabulated value of the test statistics and the computed value. If the computed value exceeds the tabulated value we reject the null hypothesis.

The sampling distribution of a test statistic under the null hypothesis is called the 'null distribution'. When the data depicts strong evidence against the null hypothesis, the value of test statistic becomes very large. By observing the computed value of the test statistic we draw inferences. Apart from the test statistic econometric software provides a *p-value*. The p-value indicates the probability of the null hypothesis being true. Thus, if we obtain a p-value of 0.04, it says the probability of the null hypothesis being true is 0.04 or 4 per cent. Therefore, if we take 5 per cent level of significance, we reject the null hypothesis.

3.3 ESTIMATION METHODS

In Unit 2 we described about two concepts; point estimation and interval estimation. We also discussed about certain probability distribution functions such as normal, *t*, *F* and chi-square.

There are basically three estimation methods: (i) least squares, (ii) maximum likelihood, and (iii) method of moments. We will use the least squares estimation method extensively in this course. In Unit 7 of this course we have introduced the maximum likelihood method. You are not introduced to 'Method of Moments' in this course.

In Unit 5 of the course BECC 107 we discussed with the concept of regression. In Section 5.9 that Unit we mentioned that the error variable in the regression should be minimised. For that purpose, we minimised the sum of squares of the error terms ($\sum u_i^2$). Now you can guess why it is called the least squares method. In this course we confine to ordinary least squares (OLS) method. We deal with OLS method first with the two-variable case. Subsequently, we extend it to more than two variables. This leads us the multiple regression model.

The name ordinary least squares (OLS) suggests that it is the simplest of the least squares methods. It implies that further complexities can be brought into the OLS method. Correctly so; there are generalised least squares (GLS), two-stage least squares (2SLS), three-stage least squares (3SLS), etc. Therefore, be careful when you read about the least squares method – notice which method the text is

referring to. When you come across the term GLS in some context do not confuse it with OLS – both methods are different. In both OLS and GLS the sum of squares of the error terms is minimised (that is why both are referred to as least squares method) but there is some transformation of the regression model in the case of GLS. The advanced methods of least squares are not dealt with in this course. Remember that for carrying out the least squares method you do not need to assume any probability distribution function about the variables.

The maximum likelihood (ML) method assumes a probability distribution about the variables. Normal distribution is the most commonly used probability distribution function in maximum likelihood estimation. In ML method we form a likelihood function, which is derived from the probability distribution function. Note that in econometrics we are given the data – the data is obtained from a sample survey. We estimate the parameters of the regression model, under that the assumption that the data follows certain probability distribution function (for example, normal distribution). The likelihood function can follow any of the probability distribution functions; not just normal distribution. Recall from your statistics course that in probability distribution function we are given the parameters and we find out the probability of occurrence of particular dataset. In ML method, we do the opposite – we are provided with the data, and we are estimating the parameters.

The method of moments (MOM) makes use of the moment generating function (MGF) properties. You have been introduced to the concept of 'moments' in Unit 4 of BECC 107. The moment generating function of certain probability distributions are used for estimation of the parameters. The method of moments is quite advanced and beyond the scope of this course.

3.4 REJECTION REGION AND TYPES OF ERRORS

In the previous Unit we discussed about point estimation and interval estimation. The underlying idea behind hypothesis testing and interval estimation is the same. Recall that a confidence interval is built around sample mean with certain confidence level. A confidence level of 95 per cent implies that in 95 per cent cases the population mean would remain in the confidence interval estimated from the sample mean. It is implicit that in 5 per cent cases the population mean will not remain within the confidence interval. Note that when the population mean does not remain within the confidence interval our test statistic should reject the null hypothesis.

3.4.1 Rejection Region for Large Samples

Let us explain the concept of critical region. Sampling distribution of sample mean ($\bar{x}$) follows normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The standard deviation of a sampling distribution is known as 'standard error'.

Thus, $\bar{x}$ can be transformed into a standard normal variable, z, so that it follows normal distribution with mean 0 and standard deviation 1.

In notations, $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and $z \sim N(0,1)$.

Recall that area under the standard normal curve gives the probability for different range of values assumed by *z*. These probabilities are presented as the area under standard normal curve.

Let us explain the concept of critical region or rejection region through the standard normal curve given in Fig. 14.1 below. When we have a confidence coefficient of 95 percent, the area covered under the standard normal curve is 95 per cent. Thus 95 per cent area under the curve is bounded by $-1.96 \le z \le 1.96$. The remaining 5 per cent area is covered by $z \le -1.96$ and $z \ge 1.96$. Thus 2.5 per cent of area on both sides of the standard normal curve constitute the rejection region. This area is shown in Fig. 3.1. If the sample mean falls in the rejection region we reject the null hypothesis.

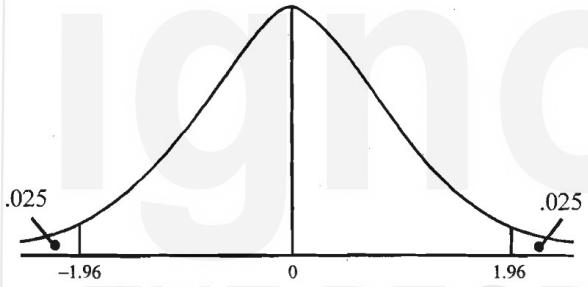


Fig. 3.1: Critical Regions

3.4.2 One-tail and Two-tail Tests

In Fig. 3.1 we have shown the rejection region on both sides of the standard normal curve. However, in many cases we may place the rejection region on one side (either left or right) of the standard normal curve. Remember that if α is the level of significance, then for a two-tail test $\dfrac{\alpha}{2}$ area is placed on both sides of the standard normal curve. But if it is a one-tail test, then α area is placed on one-side of the standard normal curve. Thus the critical value for one-tail and two tail test differ.

The selection of one-tail or two-tail test depends upon the formulation of the alternative hypothesis. When the alternative hypothesis is of the type $H_A : \bar{x} \neq \mu$ we have a two-tail test, because $\bar{x}$ could be either greater than or less than μ. On the other hand, if alternative hypothesis is of the type $H_A : \bar{x} < \mu$, then entire rejection is on the left hand side of the standard normal curve. Similarly, if the alternative hypothesis is of the type $H_A : \bar{x} > \mu$, then the entire rejection is on the right hand side of the standard normal curve.

The critical values for z depend upon the level of significance. In the appendix tables at the end of this book Table 14.1 these critical values for certain specified levels of significance (α) are given.

3.4.3 Rejection Region for Small Samples

In the case of small samples ($n \leq 30$), if population standard deviation is known we apply z-statistic for hypothesis testing. On the other hand, if population standard deviation is not known we apply t-statistic. The same criteria apply to hypothesis testing also.

In the case of small samples if population standard deviation is known the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \qquad \ldots(3.1)$$

On the other hand, if population standard deviation is not known the test statistic is

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(3.2)$$

In the case of t-distribution, however, the area under the curve (which implies probability) changes according to degrees of freedom. Thus while finding the critical value of t we should take into account the degrees of freedom. You should remember two things while finding critical value of t. These are: i) level of significance, and ii) degrees of freedom.

3.5 TYPES OF ERRORS

In hypothesis testing we reject or do not reject a hypothesis with certain degree of confidence. As you know, a confidence coefficient of 0.95 implies that in 95 out of 100 samples the parameter remains within the acceptance region and in 5 per cent cases the parameter remains in the rejection region. Thus in 5 per cent cases the sample is drawn from the population but sample mean is too far away from the population mean. In such cases the sample belongs to the population but our test procedure rejects it. Obviously we commit an error such that H_0 is true but gets rejected. This is called 'Type I error'. Similarly there could be situations when the H_0 is not true, but on the basis of sample information we do not reject it. Such an error in decision making is termed 'Type II error' (see Table 3.1).

Note that Type I error specifies how much error we are in a position to tolerate. Type I error is equal to the level of significance, and is denoted by α. Remember that confidence coefficient is equal to $1 - \alpha$.

The probability of committing a type I error is designated as α and is called the level of significance. The probability of committing type II error is called β. Thus,

Type I error = α = prob (rejecting H_0 | H_0 is true)

Type II error = β = prob (accepting H_0 | H_0 is false)

Table 3.1: Type of Errors

	H_0 **true**	H_0 **not true**
Reject H_0	Type I Error	Correct decision
Do not reject H_0	Correct decision	Type II Error

Check Your Progress 1

1) Distinguish between one-tail and two-tail tests.

.....

.....

.....

.....

.....

.....

2) Distinguish between Type I and Type II errors.

.....

.....

.....

.....

.....

.....

3) Suppose the cholesterol level of an individual is normally distributed with mean of 180 and standard deviation of 20. Cholesterol level of over 225 is diagnosed as not healthy.

   a) What is the probability of making type I error?

   b) What level should people be diagnosed as not healthy if we want the probability of type I error to be 2%?

.....

.....

.....

.....

.....

3.6 POWER OF TEST

As pointed out above, there are types I and type II errors in hypothesis testing. Thus, there are two types of risks: (i) α represents the probability that the null hypothesis is rejected when it is true and should not be rejected. (ii) β represents the probability that null hypothesis is not rejected when in reality it is false. The power of test is referred to as $(1 - \beta)$, that is the complement of β. It is basically the probability of not committing a type II error.

A 95% confidence coefficient means that we are prepared to accept at most 5% probability of committing type I error. We do not want to reject a true hypothesis by more than 5 out of 100 times. This is called 5% level of significance.

The power of test depends on the extent of difference between the actual population mean and the hypothesized mean. If the difference is large then the power of test will be much greater than if the difference is small. Therefore, selection of level of significance α is very crucial. Selecting large value of α makes it easier to reject the null hypothesis thereby increasing the power of the test $(1 - \beta)$.

At the same time increasing the sample size increases the precision in the estimates and increases the ability to detect the difference between the population parameter and sample, increasing the power of the test.

3.7 APPROACHES TO PARAMETER ESTIMATION

In statistical hypothesis testing, estimation theory deals with estimating the values of parameters based on measurement of empirical data that has a random component. The method of estimation requires setting up of a null hypothesis and a corresponding alternative hypothesis, which are further rejected or not rejected based on the two approaches used to make decision regarding the null hypothesis. The two methods have been described in the following section.

3.7.1 Test of Significance Approach

Any test statistic can be used for the test of significance approach to hypothesis testing. Let us consider the t-statistic.

$$t = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \qquad \ldots (3.3)$$

If the difference between $\bar{X}$ and μ_X is small, $|t|$ value will also be small, where $|t|$ is the absolute value of t-statistic. You should note that t = 0, if $\bar{X} = \mu_X$. In this case we do not reject the null hypothesis. As | t | gets larger, we would be more inclined to reject the null hypothesis.

Example. Suppose for a dataset $\bar{X} = 23.25, S_X = 4.49,$ and $n = 28$. Our nulland alternative hypothesis are

H_0: $\mu_X = 18.5$ and H_A: $\mu_X \neq 18.5$

$$\therefore \quad t = \frac{23.25-18.5}{9.49/\sqrt{28}} = 2.6486 \qquad \ldots (3.4)$$

We need to specify α, the probability of rejecting the null hypothesis (probability of commuting type I error). Let us fix α at 5%.

$H_0: \mu_X = 18.5$

$H_A: \mu_X \neq 18.5$ (two-tailed test)

Since the computed *t* value is 2.6486. This value lies in the right-tail critical region of the *t*-distribution. We therefore reject the null hypothesis (H_0) that the true population mean is 18.5.

A test is statistically significant means that we one can reject the null hypothesis. This implies that the probability of observed difference between the sample value and the critical value (also called tabulated value) is not small and is not due to chance.

A test is statistically not significant means that we do not reject the null hypothesis. The difference between the sample value and the critical value could be due to sampling variation or due to chance mechanism.

3.7.2 Confidence Interval Approach

Let us assume that the level of significance or the probability of commuting type I error is fixed at α = 5%. Suppose the alternative hypothesis is two-sided. Assume that we apply *t*-distribution since variance is not known. From the *t* table we find the critical value of *t* at 8 degree of freedom $(n-K) = (10-2)$ at α = 5%. We find out the value to be 2.360. Thus we construct the confidence interval

$$P(-2.360 \leq t \leq 2.306) = 0.95 \qquad \ldots (3.5)$$

The probability that *t* value lies between the limits ($-2.360 \leq t \leq 2.360$) is 0.95 or 95%. The values -2.360 and 2.360 are the critical *t* values.

If we substitute the t from equation (3.2)

$$P\left(-2.306 \leq \frac{b_2-\beta_2}{\mathrm{SE(b_2)}} \leq 2.306\right) = 0.95 \qquad \ldots (3.6)$$

As we will see in Unit 4, $\mathrm{SE(b_2)}$ is $\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}$

If we substitute the above value in equation (3.6) and re-arrange terms we obtain

$$P\left(b_2 - 2.306\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}} \leq \beta_2 \leq b_2 + 2.306\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}\right) = 0.95 \qquad \ldots (3.7)$$

Equation (3.7) provides a 95% confidence interval for the parameter β_2. Such a confidence interval is known as the region of acceptance (H_0). The area outside the confidence interval is known as the rejection region (H_A).

If the confidence interval includes the value of the parameter β_2,we do not reject the hypothesis. But if the parameter lies outside the confidence interval, we reject the null hypothesis.

Check Your Progress 2

1) What is meant by power of a test?

.....

.....

.....

.....

.....

.....

2) Explain how a confidence interval is built.

.....

.....

.....

.....

.....

.....

3.8 LET US SUM UP

This unit elaborated the procedure of statistical inference regarding the population parameters. There are two approaches to hypothesis testing of population parameters: test of significance approach, and confidence interval approach. The unit also pointed out that there are errors involved in testing of hypothesis. While making a decision regarding acceptance or rejection of a hypothesis, two types of error may be committed: type I error, and type II error. Power of a test is the probability of not committing a type II error, i.e., rejecting H_0 when it is false is $(1 - \beta)$.

3.9 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Sub-Section 3.4.2 and answer.

2) We have given the types of errors in table 3.1. You should elaborate on that.

3) a) In order to test this we use z-statistics $z = (X - \mu)/\sigma$, $z= (225 - 180)/ 20 = 2.25$

b) The area corresponding to the z value of 2.25 is 0.0122, which the probability of making type I error. An area of tail as 2% corresponds to $Z = 2.05$.

$Z = (X - \mu)/\sigma$

$2.05 = (X - \mu)/ 20$, i.e., $(X - \mu) = 2.05 * 20 = 41$

$X = 41 + 180 = 221$

Check Your Progress 2

1) Go through Section 3.6 and answer.

2) Go through Section 3.7.2 and answer.

UNIT 4 SIMPLE LINEAR REGRESSION MODEL: ESTIMATION*

Structure

4.0 OBJECTIVES

After going through this unit, you should be able to

- describe the classical linear regression model;
- differentiate between Population Regression Function (PRF) and Sample Regression Function (SRF);
- find out the Ordinary Least Squares (OLS) estimators;
- describe the properties of OLS estimators;
- explain the concept of goodness of fit of regression equation; and
- describe the coefficient of determination and its properties.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

4.1 INTRODUCTION

In Unit 5 of the course BECC 107: Statistical methods for Economics we discussed the topics correlation and regression. In that Unit we gave a brief idea about the concept of regression. You already know that there are two types of variables in regression analysis: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected *n* pairs of observations in X and Y. The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. It means that the relationship between X and Y is in the form of a straight line, and therefore, it is called linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Thus in general terms we can express the relationship between X and Y as follows in equation (4.1).

$$Y = f(X) \quad \ldots (4.1)$$

In this block (Units 4, 5 and 6) we will consider simple linear regression models with two variables only. The multiple regression model comprising more than one explanatory variable will be discussed in the next block.

Regression analysis may have the following objectives:

- To estimate the mean or average value of the dependent variable, given the values of the independent variables.
- To test the hypotheses regarding the underlying economic theory. For example, one may test the hypotheses that the price elasticity of demand is (–)1 that is, the demand is perfectly elastic, assuming other factors affecting the demand are held constant.
- To predict the mean value of the dependent variable given the values of the independent variable.

4.2 POPULATION REGRESSION FUNCTION

A population regression function hypothesizes a theoretical relationship between a dependent variable and a set of independent or explanatory variables. It is a linear function. The function defines how the conditional expectation of a variable Y responds to the changes in independent variable *X*.

$$Y_i = E(Y_i|X_i) + u_i \quad \ldots (4.2)$$

The function consists of a deterministic component $E(Y|X)$ and a non-deterministic or 'stochastic' component u, as depicted in equation (4.2).

We are concerned about examining the determinants of dependent variable (Y) conditional upon the given values of impendent variables (X).

4.2.1 Deterministic Component

The conditional expectation of Y constitutes the deterministic component of the regression model. It is obtained in the form of a deterministic line. It is also known as the Population Regression Line (PRL). The non-deterministic or stochastic component is represented by a random error term, denoted by u_i.

Let us take an example. Suppose we want to examine the impact of weekly personal disposable income (PDI) on the weekly expenditure for a set of population, then we consider weekly PDI as the independent variable (Y) and weekly expenditure as the dependent variable (X). For each given value of weekly PDI, the average value of weekly expenditure is plotted on the vertical axis. People with higher income are likely to spend more, therefore intuitively, the relationship between weekly PDI and weekly expenditure is positive. Thus the following Population Regression Line is obtained and plotted on a graph as explained below.

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i \quad \ldots (4.3)$$

Note that in equation (4.3), β_1 and β_2 are the parameters. Here β_1 is the intercept of the population regression function. It indicates the expected value of the dependent variable when the explanatory variable is zero. Further, β_2 is the slope of the population regression function. It indicates the magnitude by which the dependent variable will change if there is a one unit change in the independent variable. The population parameters describe the relationship between the dependent variable and the independent variable in the population.

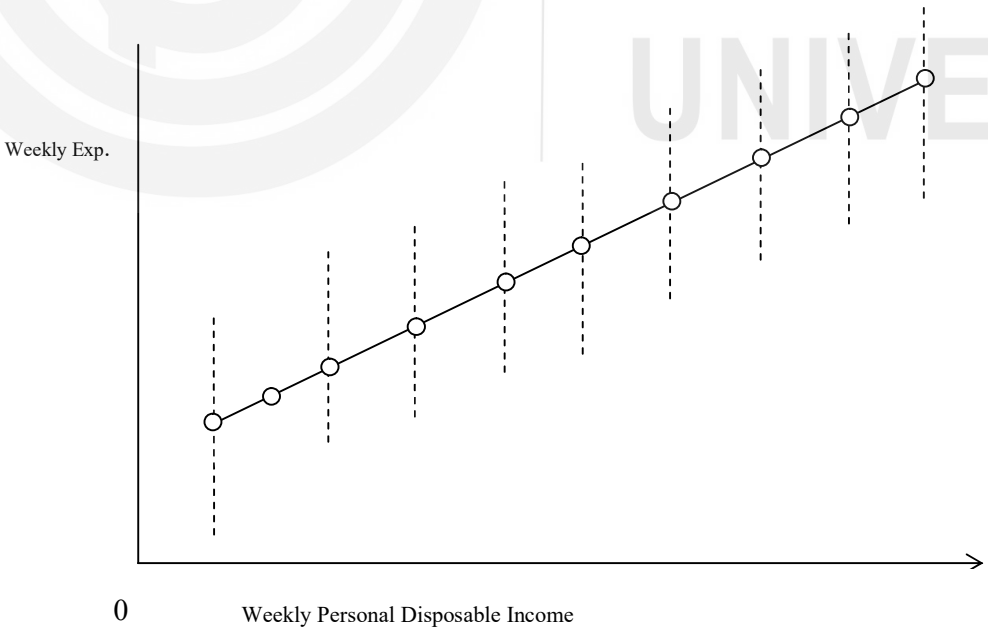


Fig. 4.1: Weekly Personal Disposable Income

Look into the circled points in Fig. 4.1. These points represent the mean or the average value of Y corresponding to various X_i. They are called the conditional means or conditional expectation values. If we connect the various expected values of Y, the resulting lines is called the Population Regression Line (PRL).

4.2.2 Stochastic Component

When we collect data from a sample, we do not a deterministic relationship between X and Y. For example, for the same level of income the expenditure of two persons could be different. Suppose there are two persons with monthly income of Rs. 20000 per month. While the monthly expenditure of one person is Rs. 15000, that of the other person could be Rs. 19000. The differences in monthly expenditure for the second person could be higher due to his health condition or living style. Such differences in the dependent variable are captured by the stochastic error term. In Fig. 4.1, for a particular value of X, the value of the Y variable is depicted by a vertical dotted line. The expected value of Y for a particular value of X is circled (see Fig. 4.1).

Thus, there is a need to specify the stochastic relationship between X and Y. The specification of the sample regression function (SRF) is

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots (4.5)$$

In equation (4.5) the term u_i is called stochastic error or random error.

The first component of equation (4.5) is the deterministic component ($\beta_1 + \beta 2Xi$, which we have already discussed. The deterministic component is the mean or average expenditure in the example under consideration. The deterministic component is also called the systematic or deterministic component.

The second component u_i is called the random component (determined non-systematically by factors other than income). The error term u_i is also known as the 'noise component'. The error term u_i is a random variable. The value of u_i cannot be controlled or known.

There are three reasons for including the error term u_i in a regression model: (i) The error term represents the influence of those variables that are not explicitly introduced in the regression model. For example, there are several variables that influence consumption expenditure of a household (such as number of family members, health status, neighbourhood, etc.). These variables affect the dependent variable, and there exists intrinsic randomness between X and Y. (ii) Human behaviour is not predictable. This sort of randomness is reflected and captured by the random error term. (iii) The errors in measuring data such as rounding off of annual family income, absence of many students from the school, etc.

Because of the randomness the actual value of the data would either remain above or below the expected value of the dependent variable. In other words, the actual value will deviate from the average, that is, the systematic component.

Having understood the elementary concept of Population Regression Function and Population Regression Line (PRL), the following section describes the estimation of PRL using the sample.

Check Your Progress 1

1) What are the objectives of estimating regression models?

.....

.....

.....

.....

.....

.....

2) Why does the average value of the dependent variable differ from the actual value?

.....

.....

.....

.....

.....

3) Why do we include an error term ($\boldsymbol{u_i}$) to the regression model?

.....

.....

.....

.....

.....

4.3 SAMPLE REGRESSION FUNCTION

We rarely have the data related to the entire population at our disposal. We only have a sample from the population. Thus, we need to use the sample to estimate the population parameters. We may not be able to find out the population regression line (PRL) because of sampling fluctuations or sampling error. Suppose we have two samples from the given population. Using the samples separately, we obtain Sample Regression Lines (SRLs). A sample represents the population. In Fig. 4.2 we have shown two sample regression lines, SRL_1 and SRL_2.

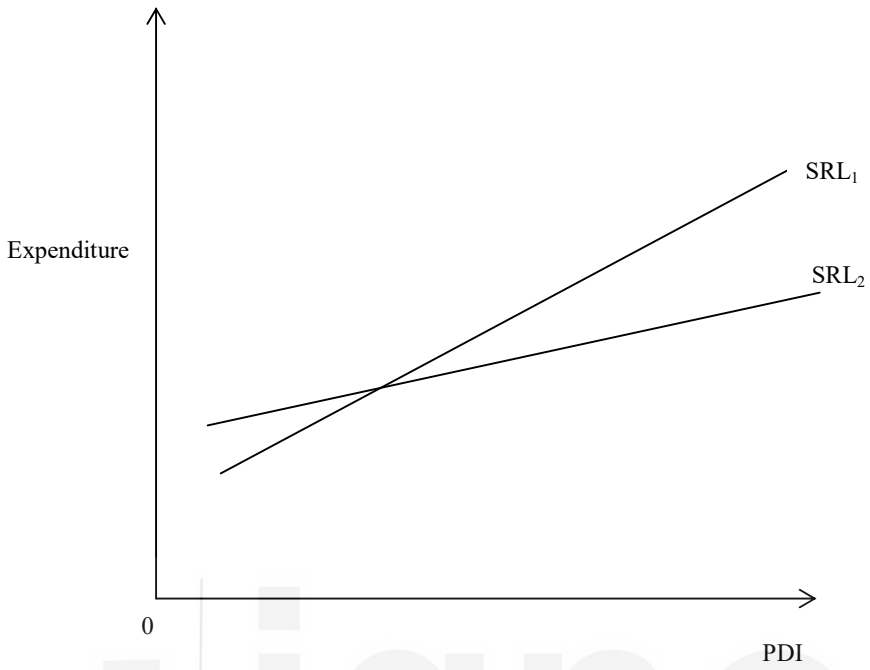


Fig: 4.2: Two Sample Regression Lines

Both the sample regression lines represent the population regression line. However, due to sampling fluctuation, the slope and intercept of both the SRLs are different. Analogous to population regression function (PRF) that underlies the PRL, we develop the concept of Sample Regression Function (SRF) comprising Sample Regression Line (SRL) and the error term u_i.

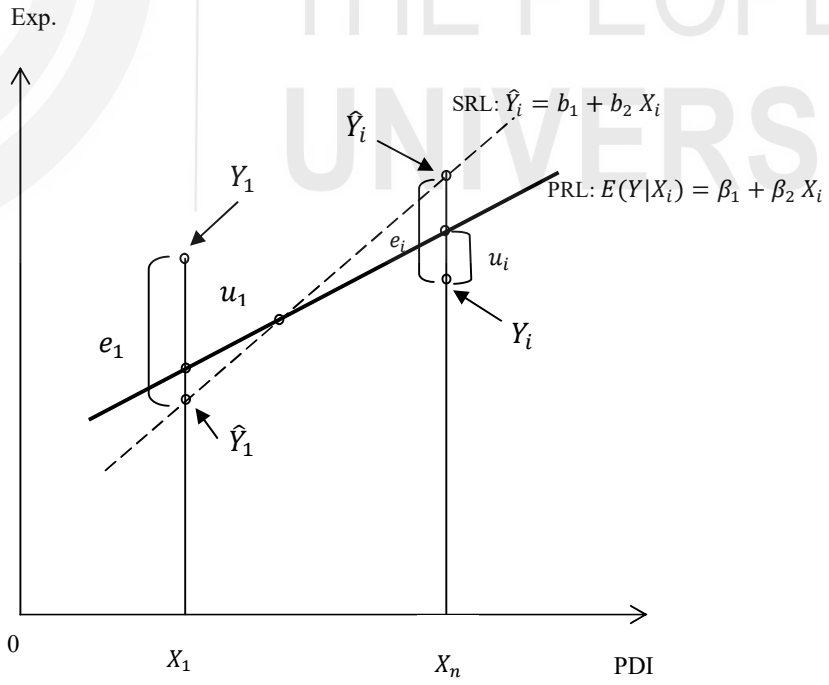


Fig 4.3: Population Regression Line and Sample Regression Line

In Fig. 4.3 we depict the population regression line (PRL) and the sample regression line (SRL). We observe that the slopes of both the lines are different. Thus, $b_1 \neq \beta_1$ and $b_2 \neq \beta_2$. Let us consider a particular value of the explanatory variable, X_1. The corresponding value of the explained variable is Y_1. On the basis of the sample regression line we obtain estimated value of the explained variable, $\hat{Y}_1$. Now let us find out the distinction between the error term (u) and the residual (e). The distance between the actual value Y_1 and the corresponding point on the population regression line is u_1. This error u_1 is not known to us, because we do not know the values of β_1 and β_2. What we know is $\hat{Y}_1$, which is estimated on the basis of b_1 and b_2. The distance between Y_1 and $\hat{Y}_1$ is the residual, e_1.

The population regression line as given in equation (4.2) is

$Y_i = E(Y_i|X_i) + u_i$

The sample regression line that we estimate is given by

$\hat{Y}_i = b_1 + b_2 X_i$ … (4.6)

In equation (4.6) the symbol (^) is read as 'hat' or 'cap'. Thus, $\hat{Y}_i$ is read as 'Y_i-hat'.

You should remember that what we observe are proxies b_1, b_2 and e in place of β_1, β_2 and u_i.

$Y_i = \hat{Y}_i + e_i = b_1 + b_2 X_i + e_i$ ... (4.7)

where $\hat{Y}_i$ = estimator of $E(Y|X_i)$, the estimator of the population conditional mean $\hat{Y}_i$ is an estimator (or a sample statistic) in equation (4.7). A particular value obtained by the estimator is considered an estimate.

The actual value of Y is obtained by adding the residual term to the estimated value of Y, also referred as the residual. The residual is the estimated value of random error term of the population regression function. The sample regression function in equation (4.7) is combination of sample regression line given by $\hat{Y}_i$ and the estimated residual term e_i. The dark straight line in Fig. 4.3 is the Population Regression Line (PRL) and it is given by the following equation:

$E(Y|X) = \beta_1 + \beta_2 X_i$. …(4.8)

Therefore, the Population Regression function (PRF) can be expressed as

$Y_i = E(Y_i|X_i) + u_i$

Or,

$Y_i = \beta_1 + \beta_2 X_i + u_i$ … (4.9)

Thus, the Population Regression Function in equation (4.9) is a combination of population regression line (PRL) $E(Y_i|X_i)$ and random error term u_i. The SRF is only an approximation of PRF. We attempt to find the most appropriate sample that yields estimators b_1 and b_2 which are as close as possible to population

parameters $\beta_1$and β_2. In other words, $b_1$is as close as possible to β_1, and b_2 is as close as possible to β_2.

4.4 ASSUMPTIONS OF CLASSICAL REGRESSION MODEL

A linear regression model is based on certain assumptions as specified below. If a regression model fulfils the following assumptions, it is called the classical linear regression model (CLRM). The assumptions of CLRM are as follows:

(i) The regression model is linear in parameters. It may or may not be linear in variables. For example, the equation given below is linear in parameters as well as variables as shown in equation (4.10)

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \ldots(4.10)$$

(ii) The explanatory variable is not correlated with the disturbance term *u.* This assumption requires that $\sum u_i X_i = 0$. In other words, the covariance between error term and explanatory variable is zero. This assumption is automatically fulfilled if X is non-stochastic. It requires that the X_i values are kept fixed in repeated samples.

(iii) The expected value or mean value of the error term u is zero. In symbols, $E(u_i|X_i) = 0$. It does not mean that all error terms are zero. It implies that the error terms cancel out each other.

(iv) The variance of each u_i is constant. In symbols, $var(u_i) = \sigma^2$. The conditional distribution of the error term has been displayed in Fig. 4.4(a). The corresponding error variance for a specific value of the error term has been depicted in Fig. 4.4(b). From the figure you can make out that the error variance is constant at all levels of the X variable. It describes the case of 'homoscedasticity'.

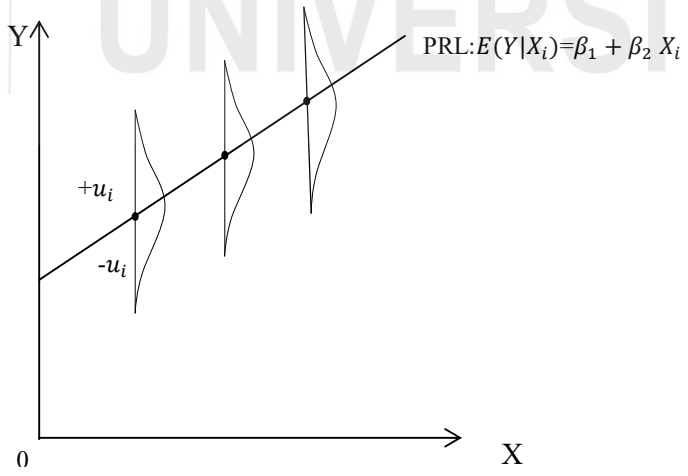


Fig 4.4 (a) Conditional Distribution of Error Term u_i

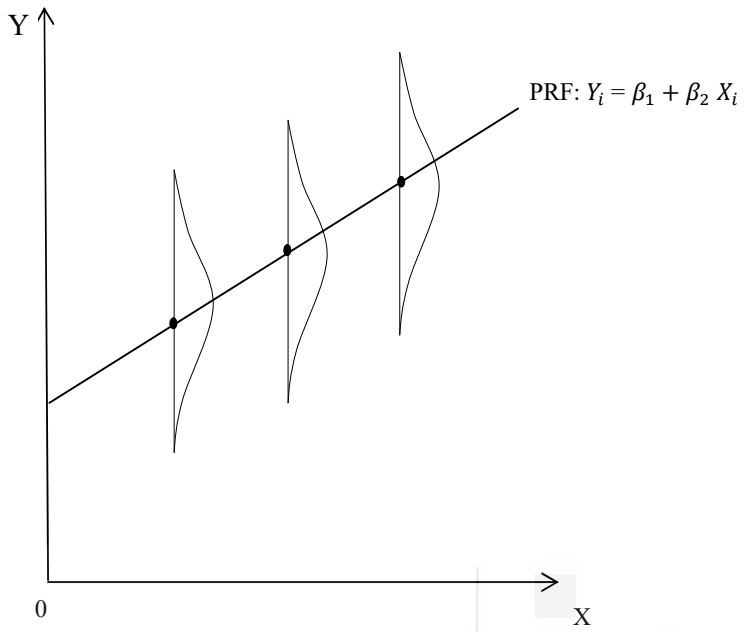


Fig 4.4 (b) Homoscedasticity (equal variance)

Fig. 4.5 depicts the case of unequal error variance, i.e., heteroscedasticity. Here the variance of the error terms varies across the values of X_i.

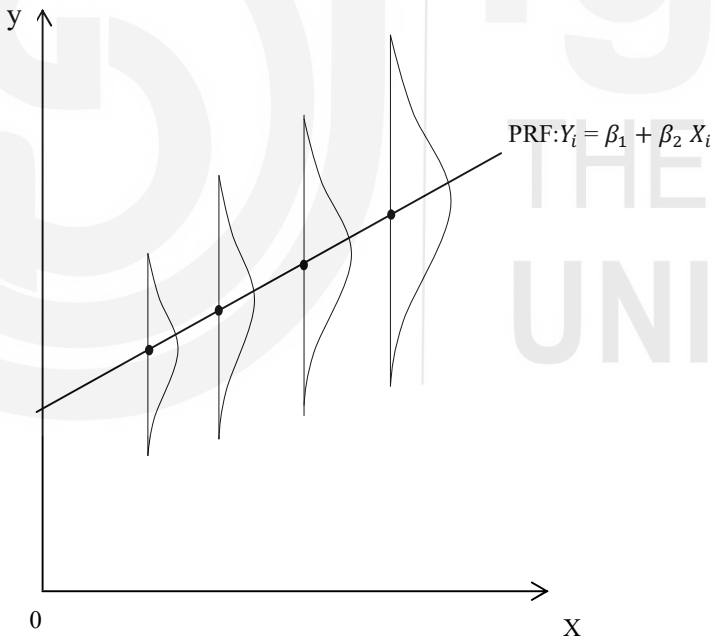


Fig. 4.5: Case of Heteroscedasticity (Unequal Variance)

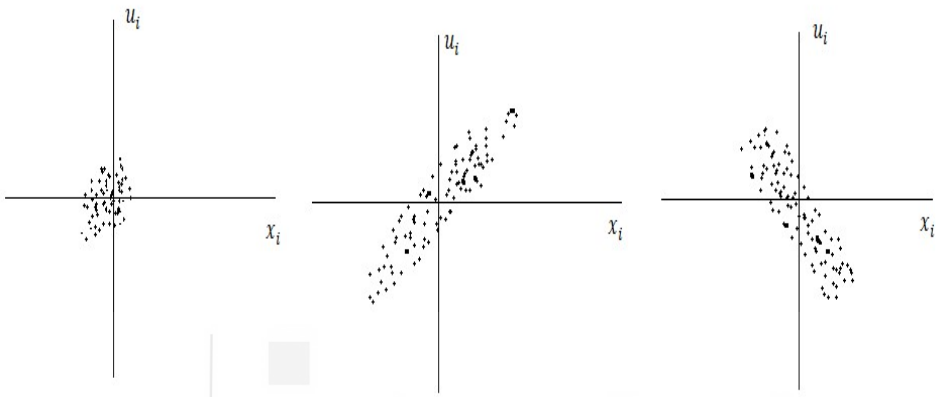
(v) There is no correlation between the two error terms. This is the assumption of no autocorrelation.

$$cov(u_i, u_j) = 0 \quad i \neq j$$

It implies that there is no systematic relationship between two error terms. This assumption implies that the error terms u_i are random.

Since two error terms are assumed to be uncorrelated, any two Y values will also be uncorrelated, i.e., $cov(Y_i, Y_j) = 0$.

Fig 4.6(i) depicts the case of no autocorrelation. Fig 4.6(ii) depicts positive autocorrelation, and Fig 4.7(iii) shows the case of negative autocorrelation.



(i) No Autocorrelation *(ii) Positive Autocorrelation* *(iii) Negative Autocorrelation*

Fig 4.6: Various Cases of Autocorrelation

(vi) The regression model is correctly specified, that is, there is no specification error in the model. If certain relevant variable is not included or certain irrelevant variable is included in the regression model then we commit model specification error. For instance, suppose we study the demand for automobiles. If we take the price of automobiles only and do not include the income of the consumer income then there is some specification error. Similarly, if we do not take into account costs of adverting, financing, gasoline prices, etc., we will be committing model specification error (we will discuss the issue of specification error in Unit 13).

4.5 ORDINARY LEAST SQUARES METHOD OF ESTIMATION

As mentioned in Unit 1 of this course, we need to estimate the parameters of the regression model. There are quite a few methods of estimation of the parameters. In this course will discuss about two such methods: (i) Least Squares, and (ii) Maximum Likelihood. We discuss about the Ordinary Least Squares (OLS) method below.

The Ordinary Least Squares (OLS) method estimates the parameters of a linear regression model by minimising the error sum of squares (ESS). In other words, it minimizes the sum of the squares of the differences between the observed dependent variable (Y_i) and the predicted or expected value of the dependent variable ($\hat{Y}_i$).

In symbols,

$e_i = (Y_i - \hat{Y}_i)$

$e_i^2 = (Y_i - \hat{Y}_i)^2$

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ ... (4.11)

In OLS method we minimise $\sum_{i=1}^{n} e_i^2$.

We know that

$\hat{Y}_i = b_1 + b_2 X_i$

If we substitute the value of $\hat{Y}_i$ in equation (4.11) we obtain

$\sum_{i=1}^{n} e_i^2 = \Sigma(Y_i - b_1 - b_2 X_i)^2$

The first order condition of minimization requires that the partial derivatives are equal to zero. Note that we have to decide on the values of b_1 and b_2 such that ESS is the minimum. Thus, we have take partial derivates with respect to b_1 and b_2. This implies that

$\frac{\partial \Sigma e_i^2}{\partial b_1} = 0$ … (4.13)

and

$\frac{\partial \Sigma e_i^2}{\partial b_2} = 0$ … (4.14)

From equation (4.13) we have

$2\Sigma(Y_i - b_1 - b_2 X_i)(-1) = 0$

By re-arranging terms in the above equation we obtain

$\Sigma Y_i = nb_1 + b_2 \Sigma X_i$ … (4.15)

In equation (4.15), note that *n* is the sample size.

From equation (4.14) we have

$2\Sigma(Y_i - b_1 - b_2 X_i)(-X_i) = 0$

By re-arranging terms in the above equation we obtain

$\Sigma X_i Y_i = b_1 \Sigma X_1 + b_2 \Sigma X_i^2$ … (4.16)

Equations (4.15) and (4.16) are called normal equations. We have two equations with two unknowns (b_1 and b_2) .

Thus, by solving these two normal equations we can find out unique values of b_1 and b_2.

By solving the normal equations (4.15) and (4.16) we find that

$$b_1 = \bar{Y} - b_2\bar{X} \quad \ldots (4.17)$$

and

$$b_2 = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - X)^2}$$

Let us take the variables X and Y in deviation forms such that

$$x_i = X_i - \bar{X} \qquad y_i = Y_i - \bar{Y}$$

Thus,

$$b_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad \ldots (4.18)$$

As you can see from the formula for b_2, it is simpler to write the estimator of the slope coefficient in deviation form. Expressing the values of a variable from its mean value does not change the ranking of the values, since we are subtracting the same constant from each value. It is crucial to note that b_1 and b_2 are expressed in terms of quantities computed from the sample, given by the formula in expressions in (4.17) and (4.18).

We mention below the formulae for variance and standard deviation of the estimators b_1 and b_2

$$Var(b_1) = \sigma_{b_1}^2 = \frac{\Sigma X_i^2}{n\Sigma x_i^2}\sigma^2 \quad \ldots (4.19)$$

$$SE(b_1) = \sqrt{Var(b_1)} \quad \ldots (4.20)$$

$$Var(b_2) = \sigma_{b_2}^2 = \frac{\sigma^2}{\Sigma x_i^2}$$

$$SE(b_2) = \sqrt{\text{var}}(b_2) \quad \ldots (4.21)$$

$$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n-2} = \frac{RSS}{n-2} = \frac{RSS}{d.f.} \quad \ldots (4.22)$$

$$\text{S.E. of the residual } (e_i) = \sqrt{\hat{\sigma}^2} \quad \ldots (4.23)$$

The formulae mentioned in equations (4.19), (4.20), (4.21), (4.22) and (4.23) are the variance and standard errors of estimated parameters b_1 and b_2.

Smaller the value of $\hat{\sigma}^2$, closer is the actual Y value to its estimated value. Recall that any linear function of a normally distributed variable to itself normally distributed. If b_1 and b_2 are linear functions of normally distributed variable u_i they themselves are normally distributed. Thus,

$$b_1 \sim N(\beta_1, \sigma_{b_1}^2) \quad \ldots (4.24)$$

$$b_2 \sim N(\beta_2, \sigma_{b_2}^2) \quad \ldots (4.25)$$

Check Your Progress 2

1) Distinguish between the error term and the residual by using appropriate diagram.

.....
.....
.....
.....
.....

2) Prove that the sample regression line passes through the mean values of X and Y.

.....
.....
.....
.....
.....

4.6 ALGEBRAIC PROPERTIES OF OLS ESTIMATORS

The OLS estimators b_1 and b_2 fulfil certain important properties.

a) SRF obtained by OLS method passes through sample mean values of X and Y. This mainly implies that the point $(\bar{X}, \bar{Y})$ passes through the Sample Regression Line.

$$\bar{Y} = b_1 + b_2\bar{X} \quad \ldots(4.26)$$

Mean value of residuals $\bar{e}$ is always zero $\bar{e} = \frac{\Sigma e_i}{n} = 0$. This implies that on an average, the positive and negative residual terms cancel each other.

b) $\Sigma e_i X_i = 0$ …(4.27)

The sum of product of residuals e_i and the values of explanatory variable X is zero, i.e., the two variables are uncorrelated.

c) $\Sigma e_i \hat{Y}_i = 0$ …(4.28)

The sum of product of residuals $e_i$and estimated $\hat{Y}_i$ is zero, i.e., $e_i\hat{Y}_i = 0$.

4.7 COEFFICIENT OF DETERMINATION

Let us consider the regression model:

$Y_i = \beta_1 + \beta_2 X_i + u_i$

Recall from equation (4.7) that

$Y_i = \hat{Y}_i + e_i$

If we subtract $\bar{Y}$ from both sides of the above equation, we obtain

$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \quad \ldots (4.29)$$

[Since $e_i = Y_i - \hat{Y}_i$]

In equation (4.20) there are three terms: (i) $(Y_i - \bar{Y})$ which is the variation in Y_i, (ii) $(\hat{Y}_i - \bar{Y})$ which is the explained variation, and (iii) $(Y_i - \hat{Y}_i)$ which is the unexplained or residual variation.

Now, let us use the lower case letters to indicate deviation from mean of a variable. Equation (4.30) can be written as

$$y_i = \hat{y}_i + e_i \quad \ldots (4.30)$$

Since $\sum e_i = 0$, we have $\bar{e} = 0$.

Therefore, we have $\bar{Y} = \bar{\hat{Y}}$, that is, the mean values of the actual Y and the estimated Y are the same.

Recall that

$$Y_i = b_1 + b_2 X_i + e_i \quad \ldots (4.7)$$

and

$$\bar{Y} = b_1 + b_2 \bar{X} \quad \ldots (4.26)$$

If we subtract equation (4.26) from equation (4.7), we get

$$y_i = b_2 x_i + e_i \quad \ldots (4.31)$$

If find OLS estimator of (4.31), we obtain

$$\hat{y}_i = b_2 x_i .$$

Therefore,

$$y_i = \hat{y}_i + e_i \quad \ldots (4.32)$$

Now let us takes squares of equation (4.32) on both sides and sum it over the sample. After re-arranging terms, we obtain

$$\Sigma y_i^2 = \Sigma \hat{y}_i^2 + \Sigma e_i^2 \quad \ldots (4.33)$$

Or, equivalently,

$$\Sigma y_i^2 = b_2^2 \Sigma x_i^2 + \Sigma e_i^2 \quad \ldots (4.34)$$

Equation (4.34) can be expressed in the following manner;

TSS = ESS + RSS … (4.35)

where TSS = Total Sum of Squares

ESS = Explained Sum of Squares

RSS = Residual Sum of Squares

Let us divide equation (4.35) by TSS. This gives us

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \quad \ldots (4.36)$$

Now, let us define

$$R^2 = \frac{ESS}{TSS} \quad \ldots (4.37)$$

The R^2 is called the coefficient of determination. It is considered as measure of goodness of fit of a regression model. It is an overall 'goodness of fit' that tells us how well the estimated regression line fits the actual Y values.

4.7.1 Formula of Computing R^2

Using the definition of R^2 given at equation (4.37), we can write equation (4.36) as:

$$1 = R^2 + \frac{RSS}{TSS} = R^2 + \frac{\Sigma e_i^2}{\Sigma y_i^2}$$

Therefore,

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma y_i^2} \quad \ldots (4.38)$$

You should note that R^2 gives the percentage of TSS explained by ESS. Thus, if $R^2 = 0.75$, we can say that 75 per cent variation in the dependent variable is explained by explanatory variable in the regression model. The value of R^2 or coefficient of determination lies between 0 and 1. This is mainly because it represents the ratio of explained sum of squares to total sum of squares.

Now let us look into the algebraic properties of R^2 and interpret it. When $R^2 = 0$ we have ESS $= 1$. It indicates that no proportion of the variation in the dependent variable is explained by ESS. If $R^2 = 1$, the sample regression is a perfect fit. If $R^2 = 1$, all the observations lie on the estimated regression line. A higher value of the R^2 implies a better fit of a regression model.

4.7.2 F-Statistic for Goodness of Fit

The statistical significance of a regression model is tested by the F-statistic. By using the t-test we can test the statistical significance of a particular parameter of the regression model. For example, the null hypothesis $H_0: \beta_2 = 0$ implies that there is no relationship between Y and X in the population. By using F-statistic, we can test the null hypothesis that all the parameters in the model are zero. Therefore, we use F-statistics for goodness of fit.

F-statistics for goodness of fit is given by the following:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \quad \ldots (4.39)$$

where k is the number of parameters in regression equation and n is the sample size.

4.7.3 Relationship between F and R^2

From equation (4.39) we know that $F = \frac{ESS/(k-1)}{RSS/(n-k)}$. If we divide the numerator and the denominator by TSS, we have

$$F = \frac{ESS/TSS/(k-1)}{RSS/TSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \quad \ldots (4.40)$$

Note that the F-statistic is an increasing function of R^2. An increase in the value of R^2 means an increase in the numerator and a decrease in the denominator. Now let us explain the interpretation of F-static obtained in equation (4.41). The value obtained by applying equation (4.41) to a dataset is the calculated value of F or F-calculated. We compare this value with the tabulated value or critical value of F given at the end of the book. For comparison purpose the degrees of freedom are $((k-1),(n-k))$.

If F-calculated is greater than F-critical we reject the null hypothesis $H_0: \beta_2 = 0$. An implication of the above is that the independent variables explain the dependent variable. In other words, there exists a statistically significant relationship between Y and X.

If F-calculated is less than F-critical we do not reject the null hypothesis $H_0: \beta_2 = 0$. Thus there is no significant relationship between Y and X.

4.7.4 Relationship between F and t^2

There is relationship between the F-statistic and the t-statistic in a regression model. Suppose, the number of explanatory variables $k = 2$.

$$F = \frac{ESS/(k-1)}{RSS/(n-2)}$$

For the two-variable model,

$$F = \frac{ESS/(2-1)}{RSS/(n-2)} = \frac{ESS}{RSS/(n-2)} \quad \ldots (4.41)$$

We know that ESS $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ and $RSS = \sum_{i=1}^{n} e_i^2$

Therefore,

$$\mathrm{F} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} e_i^2/(n-2)} = \frac{\sum_{i=1}^{n}([b_1 + b_2 X_i] - [b_1 + b_2 \bar{X}])^2}{\hat{\sigma}^2} \quad \ldots (4.42)$$

$$\text{Estimation of error variance} = \hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\Sigma e_i^2}{n-2} \quad \ldots (4.43)$$

$$F = \frac{1}{\hat{\sigma}^2} \cdot \sum_{i=1}^{n} b_2^2 (X_i - \bar{X})^2 \quad \ldots (4.44)$$

We know that

$$var(b_2) = \frac{\hat{\sigma}^2}{\Sigma x_i^2}$$

Substituting equation (4.43) in equation (4.44) we get,

$$F = \frac{b_2^2}{\hat{\sigma}^2} = \frac{b_2^2}{var(b_2)} = \frac{b_2^2}{[SE(b_2)]^2} = t^2 \quad \ldots (4.45)$$

Therefore, the F-statistic is equal to square of the t-statistic ($F = t^2$). The above result, however, is true for the two-variable model only. If the number of explanatory variable increases in a regression model, the above result may not hold.

Check Your Progress 3

1) Is it possible to carry out F-test on the basis of the coefficient of determination? Explain how.

.....
.....
.....
.....
.....

2) Can the coefficient of determination be greater than 1? Explain why.

.....
.....
.....
.....
.....

4.8 LET US SUM UP

In this unit we discussed about the classical linear regression model, which is based on certain assumptions. We distinguished between the population regression function and the sample regression function. We explained why a stochastic error term is added in a regression equation. We explained the meaning of each of the assumptions of the classical regression model. The procedure of obtaining OLS estimators of a regression model is given in the Unit. The unit further elaborated on the notion of goodness of fit and concept of R-squared.

4.9 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) The objectives of carrying out a regression model could be as follows:

- To estimate the mean or the average value of the dependent variable, given the values of independent variables.
- To test the hypotheses regarding the underlying economic theory. For example, one may test the hypotheses that price elasticity of demand is (–)1.
- To predict or forecast the mean value of the dependent variable given the value of the independent variable.

2) The relationship between Y and X is stochastic in nature. There is an error term added to the regression equation. The inclusion of the random error term leads to a difference between the expected value and the actual value of the dependent variable.

3) There are three reasons for inclusion of the error term in the regression model. See Sub-Section 4.2.2 for details.

Check Your Progress 2

1) Go through Section 4.3. You should explain the difference between the error term and the residual by using Fig. 4.3.

2) In the OLS method we minimise Σe_i^2 by equating its partial derivates to zero. The condition $\frac{\partial \Sigma \ e_i^2}{\partial b_1} = 0$ gives us the first normal equation:
$Y_i = nb_1 + b_2 \Sigma X_i$. If we divide this equation by the sample size, n, we obtain $\bar{Y} = b_1 + b_2 \bar{X}$. Thus, the estimated regression passes through the point $\bar{X}, \bar{Y}$.

Check Your Progress 3

1) Yes, we can carry out F-test on the basis of the R^2 value. Go through equation (4.40).

2) The value of R^2 or the coefficient of determination lies between 0 and 1. This is mainly because it represents the ratio of ESS to TSS. It indicates the proportion of variation in Y that has been explained by the explanatory variables. The numerator ESS cannot be more than the TSS. Therefore, R^2 cannot be greater than 1.

UNIT 5 SIMPLE REGRESSION MODEL: INFERENCE*

Structure

5.0 OBJECTIVES

After reading this unit, you will be able to:

- explain the concept of Testing of Hypothesis;
- derive the confidence interval for the slope coefficient in a simple linear regression model;
- explain the approach of 'test of significance' for testing the hypothesis on the estimated slope coefficient;
- describe the concept of Analysis of Variance (ANOVA);
- state the Gauss Markov Theorem with its properties; and
- derive the confidence interval for the predicted value of Y in a simple regression model.

5.1 INTRODUCTION

In Unit 4 we discussed the procedure of estimation of the values of the parameters. In this unit, we focus upon how to make inferences based on the estimates of parameters obtained. We consider a simple linear regression model with only one independent variable. This means we have one slope coefficient associated with the independent variable and one intercept term. We begin by recapitulating the basics of 'hypothesis testing'.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

5.2 TESTING OF HYPOTHESIS

Testing of hypothesis refers to assessing whether the observation or findings are compatible with the stated hypothesis or not. The word compatibility implies "sufficiently close" to the hypothesized value. It further indicates that we do not reject the stated hypothesis. The stated hypothesis is also referred to as 'Null Hypothesis' and it is denoted by H_0. The null hypothesis is usually tested against the 'alternative hypothesis', also known as maintained hypothesis. The alternative hypothesis is denoted by H_1. For instance, suppose the given population regression function is given by the equation:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{... (5.1)}$$

where X_i is personal disposable income (PDI) and Y_i is expenditure. Now, the null hypothesis is:

$$H_0 : \beta_2 = 0 \quad \text{... (5.2)}$$

while the alternative hypothesis is:

$$H_1 : \beta_2 \neq 0 \quad \text{... (5.3)}$$

We deliberately set the null hypothesis to 'zero' in order to find out whether Y is related to X at all. If X really belongs to the model, we would fully expect to reject the zero-null hypothesis H_0 in favour of the alternatives hypothesis H_1. The alternative hypothesis implies that the slope coefficient is different from zero. It could be positive or it could be negative. Similarly, the true population intercept can be tested by setting up the null hypothesis:

$$H_1 : \beta_1 = 0 \quad \text{… (5.4)}$$

while the alternative hypothesis is:

$$H_1 : \beta_1 \neq 0 \quad \text{… (5.5)}$$

The null hypothesis states that the true population intercept is equal to zero, while the alternative hypothesis states that it is not equal to zero. In case of both the null hypotheses,, i.e., for true population parameter or slope and the intercept, the null hypothesis as stated is a 'simple hypothesis'. The alternative hypothesis is composite. It is also known as a **two-sided hypothesis.** Such a two-sided alternative hypothesis reflects the fact that we do not have a strong apriori or theoretical expectation about the direction in which the alternative hypothesis must move from the null hypothesis. However, when we have a strong apriori or theoretical expectations, based on some previous research or empirical work, then the alternative hypothesis can be one-sided or unidirectional rather than two-sided. For instance, if we are sure that the true population value of slope coefficient is positive then the best way to express the two hypotheses is

$$H_0 : \beta_2 = 0$$

$H_1: \beta_2 > 0$

Let us take an example from macroeconomics. The prevailing economic theory suggests that marginal propensity to consume is positive. This means that the slope coefficient is positive. Now, suppose that the given population regression function is estimated by using a sample regression by adopting Ordinary Least Squares estimate. Let us also suppose that the results of sample regression yield the value of estimated slope coefficient as $b_2 = 0.0814$. This numerical value will change from sample to sample. We know that β_2 follows normal distribution, i.e., $b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\Sigma x_i^2}\right)$. There are two methods of testing the null hypothesis that the true population slope coefficient is equal to zero. The next two sections of this unit describe the two methods of testing of hypothesis of regression parameters.

5.3 CONFIDENCE INTERVAL

In this section, we shall derive the confidence interval for the slope parameter in equation (5.1) above. Note that the confidence interval approach is a method of testing of hypothesis. This is because it refers to the probability that a population parameter falls within the set of critical values from the Table. We make two assumptions, viz. (i) α, the level of significance on probability of committing type I error, is fixed at 5% level and (ii) the alternative hypothesis is two sided. From the *t*-table (given at the end of the book) we find the critical value of *t* at (n – k) degrees of freedom (d.f.) at $\alpha = 5\%$ is:

$$P(-2.306 \leq t \leq 2.306) = 0.95 \qquad \ldots(5.6)$$

Substituting for 't', equation (5.6) can be re-written as:

$$P\left(2.306 \leq \frac{b_2 - \beta_2}{\hat{\sigma} / \sqrt{\Sigma x_i^2}} \leq 2.306\right) = 0.95 \qquad \ldots(5.7)$$

Hence, the probability that *t* value lies between the limits –2.306, +2.306) is 0.95 or 95%. These are the critical *t* values. Substituting the value of *t* into equation (5.6) and rearranging the terms in (5.7) we get:

$$P\left[\left(b_2 - 2.306 .SE\left(b_2\right) \leq \beta_2 \leq b_2 + 2.306\, SE\left(b_2\right)\right)\right] = 0.95$$

The above equation provides a 95% confidence interval for β_2. Such a confidence interval is known as the region of acceptance (for H_0) and the area outside the confidence interval is known as the region of rejection [for (H_0)]. If this interval includes the value of β_2 we do not reject the hypothesis; but if it lies outside the confidence interval, we reject the null hypothesis.

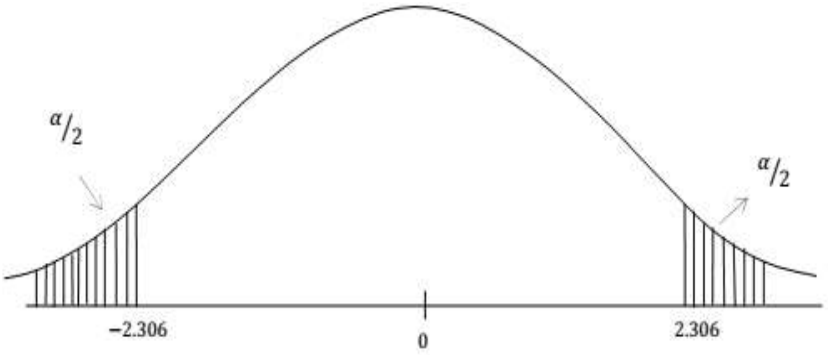


Fig 5.1: t-Distribution

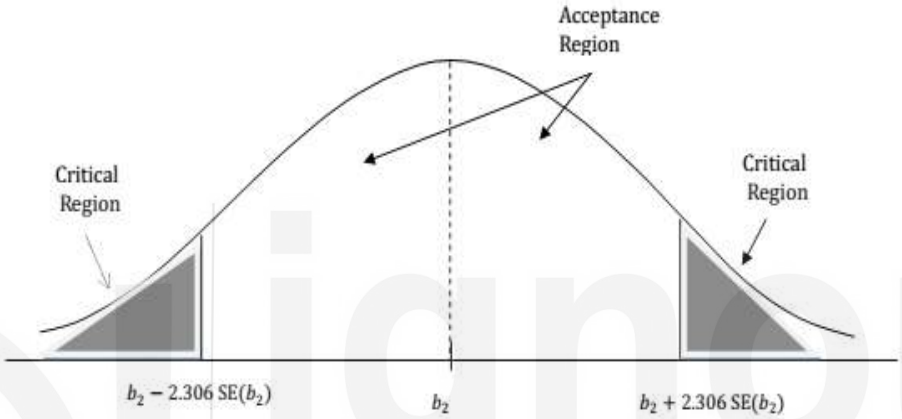


Fig 5.2: Confidence Interval for β_2

Check Your Progress 1 [answer questions in about 50-100 words]

1) State the difference between a simple and a composite hypothesis.

.....

.....

.....

.....

2) Null hypothesis is the indicator of simple and composite hypothesis. Is this statement true? Justify.

.....

.....

.....

.....

3) What is meant by a 'confidence interval'?

.....

.....

.....

.....

4) Why do we say that the interval contains the hypothesized value of true population parameter?

.....

.....

.....

.....

5.4 TEST OF SIGNIFICANCE

Test of significance approach is another method of testing of hypothesis. The decision to accept or reject H_0 is made on the basis of the value of *t*- test. It is computed by the statistic from the sample data as:

$$t = \frac{b_2 - \beta_2}{SE(b_2)} \qquad ...(5.8)$$

Equation (5.8) follows *t*-distribution with (n – k) degrees of freedom. The null hypothesis that we are testing here is:

$$H_0 : \beta_2 = \beta_2^* \qquad \ldots (5.9)$$

Note that β_2^* is some specific numerical value of β_2. Thus, the computed value of the test-statistic '*t*' will be like:

$$t = \frac{b_2 - \beta_2^*}{SE(b_2)} \qquad ...(5.10)$$

= [(estimated value) – (hypothesized value)] ÷ (standard error of estimator)

This can be computed from sample data as all values are available. The *t* value computed from (5.10) follows t distribution with ($n - k$) degrees of freedom (d.f.). This testing procedure is called the *t*-test. Fig. 5.3 depicts the region of rejection and the region of acceptance. One method of deciding on the result of the testing is to compare the computed value with the tabulated value (also called the 'critical value'). If the computed value of *t* is greater than the critical value of *t* then we reject the null hypothesis. This means we are rejecting the hypothesis that the true population parameter, or the slope coefficient, is zero. It implies that the explanatory variable plays a significant role in determining dependent variable. On the other hand, if the computed *t* value is less than critical value of *t*, then we do not reject the null hypothesis that the true value of the population parameter (or the slope coefficient) is zero. Not rejecting the null hypothesis implies that the value of slope coefficient is zero and that the explanatory variable does not play any significant role in determining the dependent variable.

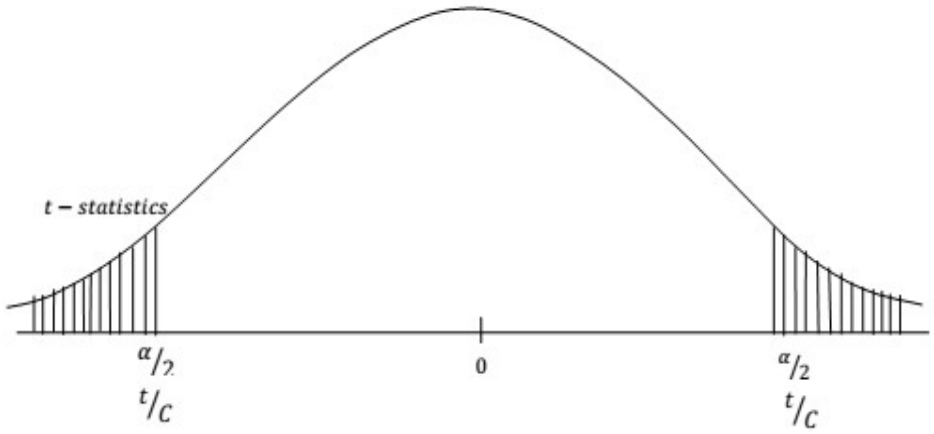


Fig 5.3: Test of Significance

In present times, when the results of the regression are obtained by computer, we usually get the *p*-value for the computed statistic. The p-value indicates the probability that the null hypothesis is true. If $p < 0.05$, we reject the null hypothesis and accept the alternative hypothesis. If $p > 0.05$, then we accept the null hypothesis. This means we base our test result at 5 percent level of significance. This also means that in 95 out of 100 independent samples, our result of the test will be similar. In other words, in 5 out of 100 cases, we could be coming to a wrong conclusion.

5.5 ANALYSIS OF VARIANCE (ANOVA)

Analysis of Variance (ANOVA) is a statistical tool used to analyse the given data for variations caused by several factors. These factors are divided into two parts: one is called the deterministic (or the systematic) part and the other is called the random part. This method of analysing the variance was developed by Ronald Fisher in 1918. Hence, this is also known as Fisher's analysis of variance. The ANOVA method separates the observed variance in the data into different components. It is used to determine the influence that the independent variables have on the dependent variable in a regression analysis. In a regression analysis ANOVA identifies the variability within a regression. Note that the total variability of dependent variable can be expressed in two parts as follows:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i) \quad \text{... (5.11)}$$

Equation (5.11) distributes the total variation in the dependent variable *Y* into two parts, i.e., the variation in mean and the residual value. Squaring each of the terms in equation (5.11) and adding over all the *n* observations, we get the following equation.

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2 \quad \text{... (5.12)}$$

The above equation can be written as: TSS = ESS + RSS where TSS is the Total Sum of Squares, ESS is the Explained Sum of Squares, and RSS is the Residual

Sum of Squares. The RSS is also called the 'Sum of Squares due to Error (SSE)'. The ratio ESS / TSS is defined as the coefficient of determination R^2. The R^2 indicates the proportion of total sum of squares explained by the regression model. An ANOVA analysis is carried out with the help of a table (Table 5.1). From such a table of analysis of variance, the *F*-statistic can be computed as: ESS/RSS. This *F*-statistic is used to test the overall level of significance of the model. The null hypothesis and the alternative hypothesis for testing the overall significance using ANOVA are given by:

H_0: Slope coefficient is zero

H_1: Slope coefficient is not equal to zero.

Table 5.1: Format of a Typical ANOVA Table

Sources	**Degrees of Freedom (df)**	**Sum of Squares**	**Mean Square**	***F* Statistics = ESS /RSS**
Model	1	$\sum(\hat{Y}_i - \bar{Y})^2$	ESS / df	
Error	*n-2*	$\sum(Y_i - \hat{Y}_i)^2$	RSS / df	
Total	*n-1*	$\sum(Y_i - \bar{Y})^2$	TSS / df	

$F = \frac{ESS/(k-1)}{RSS/(n-k)}$ gives the observed value. The *F*-critical value at (*k-1*) and (*n-k*) degrees of freedom can be located from the statistical table. When the computed *F* is > than *F*-critical, the null hypothesis is rejected. Since the alternative hypothesis is accepted, the inference is that the explanatory variable plays a crucial role in determining the dependent variable. Similarly, when the *F* computed is < than the *F*-critical, the null hypothesis is not rejected. In this case, the hypothesis that the explanatory variable plays no role in determining the dependent variable is accepted. Again, here also, we can base our inference based on the *p*-value. This means if $p < 0.05$, we reject the null hypothesis.

Check Your Progress 2 [answer questions in about 50-100 words]

1) What is ment by the 'test of significance approach' to hypothesis testing?

.....

.....

.....

.....

.....

.....

2) What does the 'level of significance' indicate?

.....

.....

.....

.....

3) What purpose does an ANOVA serve?

.....

.....

.....

.....

4) Distinguish between *t*-test in a regression model.

.....

.....

.....

.....

5.6 GAUSS-MARKOV THEOREM

This is an important theorem which gives us the condition under which the least squares estimator is the best estimator. When the assumptions of the classical linear regression model are not violated, the least-squares estimator fulfils certain optimum properties. These properties are summarised in the Gauss-Markov theorem which is stated as follows:

Gauss-Markov Theorem: Given the assumptions of classical linear regression model, the least-squares estimators, have minimum variance, in the class of all unbiased linear estimators, i.e., they are BLUE [best linear unbiased estimator(s)]. The characteristic of BLUE implies that the estimator obtained by the OLS method has the following properties.

a) It is *linear*, i.e., the estimator is a linear function of a random variable (such as the dependent variable *Y* in the regression model).
b) It is *unbiased*, i.e., its average or expected value is equal to true value [E (b_2) = β_2].
c) It has *minimum variance* in the class of all such linear unbiased estimators. In other words, such an estimator with the least variance is an efficient estimator.

Thus, in the context of regression, the OLS estimators are BLUE. This is the essence of Gauss-Markov Theorem.

5.7 PREDICTION

So far we have spoken about estimation of population parameters. In the two variable model, we derived the OLS estimators of the intercept (β_1) and slope (β_2) parameters. Prediction refers to estimation of the dependent value at a particular value of the independent variable. In other words, we use the estimated regression model to predict the value of *Y* corresponding to a given value of *X*.

Prediction is important to us for two reasons: First, it helps us in policy formulation. On the basis of the econometric model, we can find out the impact of changes in the explanatory variable on the dependent variable. Second, we can find out the robustness of our estimated model. If our econometric model is correct, the error between forecast value and actual value of the dependent variable should be small. Prediction could be of two types, as mentioned below.

5.7.1 Individual Prediction

If we predict an individual value of the dependent variable corresponding to a particular value of the explanatory variable, we obtain the individual prediction. Let us take a particular value of X, say $X = X_0$. Individual prediction of Y at $X = X_0$ in obtained by:

$$Y_0 = \beta_1 + \beta_0 X_0 + u_0 \quad \ldots (5.13)$$

We know that b_1 and b_2 are unbiased estimators of β_1 and β_2. Hence, $\hat{Y}_0$ is an unbiased predictor of $E(Y \mid X_0)$.

Therefore,

$$\hat{Y}_0 = b_1 + b_2 X_0 \quad \ldots (5.14)$$

Since $\hat{Y}_0$ is an estimator, the actual value Y_0 will be different from $\hat{Y}_0$, and there will be certain 'prediction error'.

The prediction error in $[\hat{Y}_0 - Y_0]$ is given by

$$\hat{Y}_0 - Y_0 = (b_1 + b_2 X_0) - (\beta_1 + \beta_2 X_0 + u_0) \quad \ldots (5.15)$$

We can re-arrange the terms in equation (5.15) to obtain

$$\hat{Y}_0 - Y_0 = (b_1 - \beta_1) + (b_2 - \beta_2) X_0 - u_0$$

Let us take expected value of (5.15).

$$E(\hat{Y}_0 - Y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2) X_0 - E(u_0) \quad \ldots (5.16)$$

We know that $E(b_1) = \beta_1$, $E(b_2) = \beta_2$ and $E(u_0) = 0$.

Thus, we find that expected value of prediction error is zero.

Now let us find out the variance of the prediction error.

The variance of the prediction error,

$$\mathrm{V}(\hat{Y}_0 - Y_0) = V(b_1 - \beta_1) + V(b_2 - \beta_2)X_0$$

$$+2\,X_0\,cov(b_1 - \beta_1, b_2 - \beta_2) + V(u_0) \quad \ldots (5.17)$$

We know that

$$V(b_1) = \sigma^2 \frac{\Sigma X_i^2}{n\Sigma x_i^2} \quad \ldots (5.18)$$

$$V(b_2) = \frac{\sigma^2}{\Sigma x_i^2} \quad \ldots (5.19)$$

$$Cov(b_1, b_2) = -\bar{X}\left(\frac{\sigma^2}{\Sigma x_i^2}\right) \quad \ldots (5.20)$$

By combining the above three equations and re-arranging terms, we obtain

$$V\left(\hat{Y}_0 - Y_0\right) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma x_i^2}\right] \quad \ldots (5.21)$$

Thus, Y_0 follows normal distribution with mean $\beta_1 + \beta_0 X_0$ and variance $\sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2}\right]$.

If we take estimator for σ^2, then we have

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_0 X_0)}{SE(\hat{Y}_0)} \quad \ldots (5.22)$$

On the basis of (5.22) we can construct confidence interval for $\hat{Y}_0$. Since

$t = \frac{\hat{Y}_0 - E(Y/\alpha_0)}{SE(\hat{Y}_0)}$, we have

$$P\left[-t_{\alpha/2} \le t \le t_{\alpha/2}\right] = 1 - \alpha$$

Thus, the confidence interval of $\hat{Y}_0$ is

$$P\left[(b_1 + b_2 X_0) - t_{\alpha/2} SE\left(\hat{Y}_0\right) \le (\beta_1 + \beta_2 X_0) \le (b_1 + b_2 X_0) + t_{\alpha/2} SE\left(\hat{Y}_0\right)\right] = 1 - \alpha \quad \ldots (5.23)$$

Let us lok into equation (5.21) again. We see that the variance of $\hat{Y}_0$ increases with $(X_0 - \bar{X})^2$. Thus, there is an increase in variance if X_0 is farther away from $\bar{X}$, the mean of the sample on the basis of which $b_1$and $b_2$are computed. In Fig. 5.4 we depict the confidence interval for $\hat{Y}_0$ (see the dotted line)

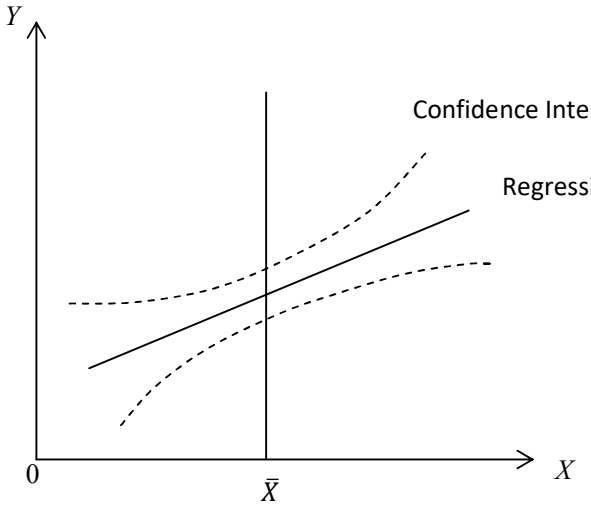


Fig. 5.4:Confidence Interval for Individual Prediction

5.7.2 Mean Prediction

It refers to prediction of expected values of Y_0, not the individual value. In other words, we are predicting the following:

$$\hat{Y}_0 = b_1 + b_2 X_0$$

Thus the error term u_0 is not added.

In the case of mean prediction, the prediction error in $\left[\hat{Y}_0 - Y_0\right]$ is given by

$$\hat{Y}_0 - Y_0 = (b_1 + b_2 X_0) - (\beta_1 + \beta_2 X_0) \quad \ldots (5.24)$$

We can re-arrange the terms in equation (5.24) to obtain

$$\hat{Y}_0 - Y_0 = (b_1 - \beta_1) + (b_2 - \beta_2) X_0$$

If we take the expected value of (5.24)

$$\mathrm{E}(\hat{Y}_0 - Y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2) X_0 \quad \ldots (5.25)$$

Thus, we find that expected value of prediction error is zero.

Now let us find out the variance of the prediction error in thecase of mean prediction.

The variance of the prediction error,

$$\mathrm{V}(\hat{Y}_0 - Y_0) = V(b_1 - \beta_1) + V(b_2 - \beta_2) X_0 + 2\, X_0\, cov(b_1 - \beta_1, b_2 - \beta_2) \quad \ldots (5.26)$$

If we compare equations (5.17) and (5.26) we notice an important change – the term $V(u_0)$ is not there in (5.26). Thus the variance of the prediction error in the case of mean prediction is less compared to individual prediction. There is a change in the variance of $\hat{Y}_0$ in the case of mean prediction, however. Variance of the prediction error, in the case of mean prediction is given by

$$V(\hat{Y}_0 - Y_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma x_i^2}\right] \quad \ldots (5.27)$$

Again, there is an increase in the variance of prediction error if X_0 is farther away from $\bar{X}$, the mean of the sample on the basis of which b_1 and b_2 are computed. It will look somewhat like the confidence interval we showed in Fig. 5.4, but the width of the confidence interval will be smaller.

An inference we draw from the above is that we can predict or forecast the value of the dependent variable, on the basis of the estimated regression equation, for a particular value of the explanatory variable (X_0). The reliability of our forecast, however, will be lesser if the particular value of X is away from $\bar{X}$.

Check Your Progress 3 [answer questions within the given space in about 50-100 words]

1) State Gauss-Markov Theorem.

.....

.....

.....

.....

.....

2) Differentiate between the two types of prediction possibilities in forecasting.

.....

.....

.....

.....

.....

5.8 LET US SUM UP

This unit explains how to make inference on the estimated results of a simple regression model. After presenting an account of hypothesis testing to recapitulate the basics, it explains the two approaches for deciding on the validation of estimated results. The two methods are: confidence interval approach and test of significance approach. The testing of overall significance of the model is explained by the technique of ANOVA. Here, the application of *F* – statistic is explained. The assumptions of classical linear regression model leads to the estimated parameters enjoying some unique properties. In light of this, the estimates are called BLUE (best linear unbiased estimates). This fact is stated in a result called the Gauss Markov theorem. The unit concludes with a detailed account of the concept of forecasting. This is once again a technique in which we have presented a confidence interval wherein the predicted or forecasted value of the dependent variable is shown to lie.

5.9 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) In case of both the null hypothesis (, i.e., for true population parameters of slope and intercept), the null hypothesis are a simple hypothesis, whereas the alternative hypothesis are composite. The former is usually equated to zero (unless equated to a known value) and the latter in stated in inequality terms. The latter is also known as **two-sided hypothesis** when stated in 'not equal to' terms**.** It is considered one sided if stated in > or < terms.

2) False. It is the alternative hypothesis that decides whether it is composite or one-sided hypothesis. If the alternative hypothesis is stated as not equal to zero then it is composite or two-tailed test. Otherwise, i.e., if the alternative hypothesis is stated in positive or negative terms, then it will be a one-sided test.

3) The confidence interval approach is a method of testing of hypothesis. It refers to the probability that a population parameter falls within the set of critical values drawn from the Table.

4) We say that the hypothesised value is contained in the interval because the value of the interval depends upon the sample or the data used for estimation. The true population parameter value is fixed but the interval changes depending on the sample.

Check Your Progress 2

1) The test of significance approach is another method of testing of hypothesis. The decision to accept or reject H_0 is made on the basis of the value of test statistic obtained from the sample data. This test statistic is given by: $t = \frac{b_2 - \beta_2}{SE(b_2)}$ and it follows $t-$ distribution with (n -1 d.f.)

2) It is a measure of the strength of evidence when the null hypothesis is rejected It concludes that the effect is statistically significant. It is the probability of rejecting the null hypothesis when it is true. This is a grave error to commit and hence is chosen in a small measure like 1% or 5%.

3) Analysis of Variance (ANOVA) is a technique or a tool used to analyse the given data in two ways or direction. One is attributed to the deterministic factors, also called the explained part or the systematic part. The other is called the random or the unexplained part. This method of analysis of variance method was developed by Ronald Fisher in 1918.

4) The *t*-test is used to test the significance of estimated individual coefficients. It is distributed as *t* with (k – 1) degrees of freedom (d.f.). where k is the number of parameters estimated including the intercept term. Thus, for a simple linear regression, it is [n – (2 – 1)] = (n –1). The *F*-distribution is used for testing the significance of the whole model. It has two parameters. The d.f. for a *F* test, in general is (k – 1) and (n – k). *K* includes the intercept term. Hence, in a simple linear regression, the d.f. for F is: (*2* – 1) and (n – 2) or 1

and (n – 2) Note that in a simple linear regression, the t test and the F test are equivalent because the number of independent variable is only one.

Check Your Progress 3

1) The Gauss-Markov theorem states that the Ordinary Least Squares (OLS) estimators are also the best linear unbiased estimator (BLUE). The presence of BLUE property implies that the estimator obtained by the OLS method retains the following properties: (i) it is linear, i.e., the estimator is a linear function of a random variable such as the dependent variable Y in the regression model; (ii) it is unbiased, i.e., its average or expected value is equal to the true value in the sense that E (b_2) = β_2; (iii) it has minimum variance in the class of all such linear unbiased estimators. Such an estimator with the least variance is also known as an efficient estimator.

2) Prediction implies predicting two types of values: prediction of conditional mean, i.e., E ($Y \mid X_0$) → a point on the population regression line. This is called as the Mean Prediction. Prediction of individual Y value, corresponding f (X_0) is called the Individual Prediction.

UNIT 6 EXTENSION OF TWO VARIABLE REGRESSION MODELS*

Structure

6.0 OBJECTIVES

After going through this Unit, you should be in a position to

- interpret regression models passing through the origin;
- explain the impact of changes in the unit of measurement of dependent and independent variables on the estimates;
- interpret parameters in semi-log and log-linear regression models; and
- identify the correct functional form of a regression model.

6.1 INTRODUCTION

In the previous two Units we have discussed how a two variable regression model can be estimated and how inferences can be drawn on the basis of the estimated regression equation. In this context we discussed about the ordinary least squares (OLS) method of estimation. Recall that the OLS estimators are the best linear unbiased estimators (BLUE) in the sense that they are the best in the class of linear regression models.

The two variable regression model has the function as follows:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \ldots (6.1)$$

*Prof. Kaustuva Barik, Indira Gandhi National Open University, New Delhi

where Y is the dependent variable and X is the independent variable. We added a stochastic error term (u_i) to the regression model. We cited three reasons for inclusion of the error term in the regression model: (i) it takes care of the excluded variables in the model, (ii) it incorporates the unpredictable human nature into the model, and (iii) it absorbs the effects measurement error, incorrect function form, etc.

We assumed that the regression model is correctly specified. All relevant variables are included in the model. No irrelevant variable is included in the regression model. In this Unit we will continue with the two variables case as in the previous two units. We also continue with the same assumptions, as mentioned in Unit 4.

Let us look into the regression model given at equation (6.1). We observe that the regression model is linear in parameters. We do not have complex forms of the parameters such as β_2^2 or $\beta_1\beta_2$ as parameters. Further, the regression model is linear variables. We do not have X^2 or $\log X$ as explanatory variable. Can we have these sorts of variables in a regression model? How do we interpret the regression model if such variables are there? We will extend the simple regression model given in equation (6.1) and explain how the interpretation of the model changes with the modifications.

6.2 REGRESSION THROUGH THE ORIGIN

Let us look into the simple regression model given at equation (6.1). There are two parameters in the regression model: β_1 and β_2. The intercept parameter is β_1 and the slope parameter is β_2. The intercept β_1 indicates the value of the dependent variable when the explanatory variable takes the value zero, i.e., $E(Y_0|X_0) = \beta_1$.

Suppose regression model takes the following form:

$$Y_i = \beta_2 X_i + u_i \quad \ldots (6.2)$$

In equation (6.2) there is only one slope parameter, β_2. There is no intercept. The implication is that the regression line passes through the origin. The population regression function is $Y = \beta_2 X_i + u_i$ and the sample regression function is $Y_i = b_2 X_i + e_i$.

Now let us apply OLS method and find out the OLS estimator b_2. As you know from Unit 4, in OLS method we minimise the error sum of squares (ESS). Thus we minimise

$$ESS = \sum e_i^2 = \sum (Y_i - b_2 X_i)^2 \quad \ldots (6.3)$$

We take derivative of the ESS and equate it to zero.

$$\frac{d\sum e_i^2}{db_2} = 0 \quad \ldots (6.4)$$

$$\frac{d\sum e_i^2}{db_2} = 2\sum (Y_i - b_2 X_i)(-X_i) = 0 \quad \ldots (6.5)$$

This implies

$$-2\sum e_i(X_iY_i - b_2X_i^2) = 0$$

$$\sum X_iY_i - b_2\sum X_i^2 = 0$$

$$b_2 = \frac{\sum X_iY_i}{\sum X_i^2} \quad \ldots (6.6)$$

The estimator given at (6.6) is unbiased. The variance of the estimator is given by

$$var(b_2) = \frac{\sigma^2}{\Sigma X_i^2} \quad \ldots (6.7)$$

Let us compare the above estimator with the estimator for the regression model $Y_i = \beta_1 + \beta_2 X_i + u_i$ (see equation (4.18) in Unit 4)

$$b_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \quad \ldots (6.8)$$

and

$$var(b_2) = \frac{\sigma^2}{\Sigma x_i^2} \quad \ldots (6.9)$$

Note that in equation (6.6) the variables are not in deviation form. Thus when we do not have an intercept in the regression model, the estimator of the slope parameter is different from that of a regression model with intercept. Both the estimators will be the same if and only if $\bar{X} = 0$.

We present a comparison between the regression model with intercept and without intercept in Table 6.1.

Table 6.1: Features of Regression Model without Intercept

Regression Model with Intercept	Regression Model without Intercept
$b_2 = \frac{\sum x_i y_i}{\sum x_i^2}$	$b_2 = \frac{\sum X_i Y_i}{\sum X_i^2}$
$\text{var}(b_2) = \frac{\sigma^2}{\sum x_i^2}$	$\text{var}(b_2) = \frac{\sigma^2}{\sum X_i^2}$
$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$	$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1}$
R^2 is non-negative	R^2 can be negative

The estimated regression model is given as

$$\hat{Y}_i = b_2 X_i \quad \ldots (6.10)$$

Note that the coefficient of determination R^2 is not appropriate for regression models without the intercept. If the intercept in a regression model is not statistically significant, then we can have regression through the origin. Otherwise, it leads to specification error. There is omission of a relevant variable.

6.3 CHANGES IN MEASUREMENT UNITS

Suppose you are given time series data on GDP and total consumption expenditure of India for 30 years. You are asked to run a regression model with consumption expenditure as dependent variable and income as the independent variable. The objective is to estimate the aggregate consumption function of India. Suppose you took GDP and Consumption Expenditure in Rs. Crore. The estimated regression equation you found is

$$Y_i = 237 + 0.65X_i \quad \ldots (6.11)$$

When you presented the results before your seniors, they pointed out that the measure of GDP and consumption expenditure should have been in Rs. Million, so that it is comprehensible outside India also. If you re-estimate the results by converting the variables, will estimates be the same? Or, do you expect some changes in the estimates? Let us discuss the issue in details.

Suppose we transform both the dependent and independent variables as follows:

$$Y_i^* = w_1 Y_i \text{ and } X_i^* = w_1 X_i \quad \ldots (6.12)$$

The regression model (6.1) can be transformed as follows:

$$Y_i^* = \beta_1 + \beta_2 X_i^* + u_i \quad \ldots (6.13)$$

Estimation of equation (6.13) by OLS method gives us the following estimators

$$b_1^* = \bar{Y}^* - b_2^* \bar{X}^* \quad \ldots (6.14)$$

$$b_2^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} \quad \ldots (6.15)$$

In a similar manner you can find out the variance of b_1^* and b_2^*, and the estimator of the error variance.

From equation (6.15) we can find out that

$$b_2^* = \frac{w_1}{w_2} b_2 \quad \ldots (6.16)$$

and

$$b_1^* = w_1 b_1 \quad \ldots (6.17)$$

Now let us look into the implications of the above.

(i) Let us begin with the dependent variable, Y_i . Suppose Y_i is doubled ($w_1 = 2$) and X_i is unchanged ($w_2 = 1$). What will happen to b_1 and b_2? Substitute the values of w_1 and w_2 in equations (6.16) and (6.17). We find that both the estimates are doubled. Thus, if the dependent variable is multiplied by a constant *c*, then all OLS coefficients will be multipled by *c*.

(ii) Now let us take the case of the independent variable. Suppose X_i is doubled ($w_2 = 2$) and Y_i is unchanged ($w_1 = 1$). On substitution of the values of w_1 and w_2 in equations (6.16) and (6.17) we find that

the slope coefficient (b_2) is halved, but the intercept (b_1) remains unchanged.

(iii) If we double both the variables X_i and Y_i, then the slope coefficient (b_2)will remain unchanged, but the intercept will change. Remember that the intercept is changed by a change in the scale of measurement of the dependent variable.

Now the question arises: Will there be a change in the t-ratio and F-value of the model? No, the *t* and *F* statistics are not affected by a change in the scale of measurement of any variable.

Check Your Progress 1

1) Under what condition should we run a regression through the origin?

.....
.....
.....
.....
.....
.....

2) What are the implications of a regression model through origin?

.....
.....
.....
.....
.....

3) What are the implications on the estimates if there is a change in the measurement scale of the explanatory variable?

.....
.....
.....
.....
.....

4) What are the implications on the estimates if there is a change in the measurement scale of the dependent variable?

.....
.....
.....
.....
.....

6.4 SEMI-LOG MODELS

In some of the cases the regression model is non-linear, but by taking logarithm on both sides of the regression equation, we get a linear model. If a model is non-linear , but becomes linear after transformation of its variables, then the model is said to be intrinsically linear . Thus, semi-log and log-linear models are intrinsically linear models. We discuss about the semi-log model in this section. We will discuss about the log-linear model in the next section.

Let us begin with a functional form as follows:

$$Y_t = e^{\beta_1+\beta_2 X_t+u_t} \quad \ldots (6.18)$$

This regression model, in its present form, is non-linear . Therefore, it cannot be estimated by OLS method. However, if we take natural logs of both the sides, we obtain

$$\ln Y_t = \beta_1 + \beta_2 t + u_t \quad \ldots (6.19)$$

It transforms into a semi-log equation. It is called a semi-log model as one of the variables is in log form.

If we take $\ln Y_t = Y_t^*$, then equation (6.19) can be written as

$$Y_t^* = \beta_1 + \beta_2 t + u_t \quad \ldots (6.20)$$

Estimation of equation (6.20) is simple. The equation is linear in parameters and in variables. Thus, we can apply OLS method to estimate the parameters. The implication of the regression model (6.20), however, is much different from the regression model (6.1).

If we take the differentiation of the regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$, we obtain

$$\frac{dY}{dt} = \beta_2 \quad \ldots (6.21)$$

Equation (6.21) shows that the slope of the regression equation is constant. An implication of the above is that the absolute change in the dependent variable for unit increase in the independent variable is constant throughout the sample. If there is an increase in X by one unit, Y increases by β_2 unit.

Now let us consider the regression model $\ln Y_t = \beta_1 + \beta_2 t + u_t$. If we take differentiation of equation (6.19) we find that

$$\frac{d\ln Y_t}{dt} = \beta_2$$

which means

$$\frac{1}{Y_t}\frac{dY_t}{dt} = \beta_2 \quad \ldots (6.22)$$

An implication of equation (6.22) is that the slope of the regression model is variable. Thus its interpretation is different from that of the regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$.

For equation (6.19), we interpret the slope coefficient (β_2) as follows: For every unit increase in *X*, there is β_2 per cent increase *Y*. Thus, for a semi-log model the change in the dependent variable in terms of percentages. The semi-log model is useful is estimating growth rates.

6.5 LOG-LINEAR MODELS

Let us consider the following regression equation:

Let us take the case of the following non-linear model

$$Y = \beta_1 X^{\beta_2} \quad \ldots (6.23)$$

This model will be intrinsically linear if it can be transformed into

$$Y^* = \beta_1 + \beta_2 X^* + u \quad \ldots (6.24)$$

Using the logarithm of each of the variable in equation (6.23), we get the following transformed equation:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad \ldots (6.25)$$

The regression model given at (6.25) is called log-linear model (because it is linear in logs of the variables) or double-log model (because both variables are in log form).

Let us take differentiation of equation (6.25) with respect to X_i

$$\frac{d(\ln Y_i)}{dX_i} = \frac{1}{Y_i} \cdot \frac{dY_i}{dX_i} \quad \ldots (6.26)$$

$$\frac{dY_i}{dX_i} = \frac{\beta_2}{X_i} \quad \ldots (6.27)$$

By combining equations (6.26) and (6.27) we find that

$$\frac{dY_i}{dX_i} = \frac{Y_i}{X_i}\beta_2$$

Or,

$$\frac{dY_i}{dX_i}\frac{X_i}{Y_i} = \beta_2 \quad \ldots (6.28)$$

A closer look at equation (6.28) shows that the slope parameter represents the elasticity between Y and X.

This attractive feature of the log-linear model has made it popular in applied work. The slope coefficient β_2 measures the elasticity of *Y* with respect to *X*, that is, the percentage change in *Y* for one per cent change in *X*. Thus, if *Y* represents the quantity of a commodity demanded and *X* its unit price, then β_2 measures the price elasticity of demand.

6.6 CHOICE OF FUNCTIONAL FORM

By you would have observed that the two variable regression model could have three functional forms as given below.

(I) $Y_i = \beta_1 + \beta_2 X_i + u_i$

(II) $lnY_i = \beta_1 + \beta_2 X_i + u_i$

(III) $\ln Y_i = \beta_1 + \beta_2 \ln x_i + u_i$

A question arises: which one is the best model? The choice of functional form depends on our objective. We should choose the model that gives us relevant answer to our queries. Suppose our objective is to estimate the impact of change in the independent variable on the dependent variable. In this case we can use model-I. On the other hand, if our objective is to estimate growth rate in the dependent variable as a result of the change in the independent variable, we should opt for semi-log model (model II). If our objective is to estimate elasticity between two variables, we choose the log-linear model.

The three regression models (Models –I, II, III) will give different estimates of the parameters. The standard error of the estimators will also be different. Further, the coefficient of determination, R^2, will be different for all three models. Can we compare the R^2 of the models and say that the model with the highest R^2 is the best fit? We cannot compare the value of R^2 obtained from regression models with different dependent variables. However, we can compare R^2 of regression models with the same dependent variable and the same estimation method. Thus the R^2 value of Model-I and Model-II cannot be compared. We can compare Model-II and Model-III in terms of their best fit.

If two regression models are almost similar in terms of their coefficient of determination, statistical significance of estimators and diagnostic checking (to be discussed in Units 13 and 14), we prefer the simpler model. The simpler model is easier to comprehend and usually accepted by others.

The log-linear regression model has certain advantages: (i) the parameters are invariant to change of scale since they measure percentage changes, (ii) the model gives elasticity figures directly, and (iii) the model moderates the problem of heteroscedasticity to some extent (see Unit 11 for the problem of heteroscedasticity).

Check Your Progress 2

1) In a semi-log model how do you interpret the slope coefficient?

.....
.....
.....

2) Describe how the slope parameter of a log-linear regression model is estimated.

.....
.....

.....
.....
.....

2) What are the advantages of the log-linear model?

.....
.....
.....
.....
.....

3) What is meant by intrinsically linear model? Can you compare the results of an intrinsically linear model with that of a linear model? Why or why not?

.....
.....
.....
.....
.....

6.7 LET US SUM UP

In this Unit we discussed about the functional forms that can be accommodated in a two variable regression model. We began with the regression model passing through the origin (there is no intercept). We pointed out the impact of changes in the scale of measurement of variables. Subsequently we considered three functional forms: the original model, the semi-log model and the log-linear model. The interpretations of the parameters in all three functional forms have been discussed in the Unit.

6.8 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) The exclusion of intercept term from a regression model has serious implication. It should be omitted only when the intercept term in the unrestricted model is statistically not significant.

2) We have listed the implications of the omission of the intercept term in table 6.1. Go through it and answer.

3) When there is a change in the measurement scale of the explanatory variable the concerned estimate is affected. If X is multiplied by *c*, the parameter is divided by *c*.

4) If Y is multiplied by c, all parameters in the model are multiplied by c.

Check Your Progress 2

1) In a semi-log model the slope parameter indicates growth rate. If there is 1 unit increase in the value of X, the expected value of Y increases by β per cent.

2) The estimation of the log-linear model is the same as the simple regression model, except that the variables are transformed. Write down the steps followed in estimation of a regression model.

3) We have mentioned three advantages in the text: (i) the parameters are invariant to change of scale since they measure percentage changes, (ii) the model gives elasticity figures directly, and (iii) the model lessens the problem of heteroscedasticity to some extent.

4) You cannot compare the results of two regression models unless the dependent variable is the same.

UNIT 7 MULTIPLE LINEAR REGRESSION MODEL: ESTIMATION*

Structure

7.0 OBJECTIVES

After going through this unit, you will be able to:

- specify the multiple regression model involving more than one explanatory variable;
- estimate the parameters of the multiple regression model by the OLS method stating their properties;
- interpret the results of an estimated multiple regression model;
- indicate the advantage of using matrix notations in multiple regression models;
- explain the maximum likelihood method of estimation showing that the 'maximum likelihood estimate (MLE)' and the OLS estimate are asymptotically similar;
- derive the expression for the coefficient of determination (R^2) for the case of a simple multiple regression model with two explanatory variables; and
- distinguish between R^2 and adjusted R^2 specifying why adjusted R^2 is preferred in practice.

7.1 INTRODUCTION

By now you are familiar with the simple regression model where there is one dependent variable and one independent variable. The dependent variable is explained by the independent variable. Now let us discuss about the multiple regression model. In a multiple regression model, there is one dependent variable

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

and more than one independent variable. The simplest possible multiple regression model is a three-variable regression model, with one dependent variable and two explanatory variables. Such a three-variable multiple regression equation or model is expressed as follows:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \ldots (7.1)$$

Throughout this unit, we shall be mostly dealing with a multiple regression model as specified in equation (7.1) above. Here, Y is the dependent variable and X_2 and X_3 are independent variables. u_i is the stochastic error term. The interpretation of this error term is the same as in the simple regression model. You may wonder as to why there is no X_1 in equation (7.1). The answer is that X_1 is implicitly taken as 1 for all observations. In the above equation, the parameter β_1 is the intercept term. We can think of Y, X_2 and X3 as some variables from economic theory. We may treat it as a demand function, where Y stands for quantity demanded of a good, and X_2 and X_3 are price of that good and the consumer's income, respectively. As another example, we can think of a production/demand function with two inputs. Here Y is the quantity produced or demanded of a good, and X_2 is the labour input, and X_3 the capital input. You can think of many similar examples.

7.2 ASSUMPTIONS OF MULTIPLE REGRESSION MODEL

Recall that the simple regression model is based on certain assumptions. These assumptions are the benchmark for a regression model. When these assumptions are fulfilled by a regression model, we call it as the classical linear regression model (CLRM). The main assumptions for the classical multiple regression models remain the same as the simple regression model. There is one change. This relates to a new assumption on multicollinearity. Since we are considering more than one independent variable X_i, it is now necessary to assume that the X_i's are not perfectly correlated. Let us recapitulate the assumptions of the CLRM with this new assumption added as follows:

(i) The regression model is linear in parameters. This assumption implies that the dependent variable is a linear function of the parameters, *βs*. The regression model could be non-linear in explanatory variables.

(ii) There is no covariance between u_i and X_i variables. This implies, in a multiple regression model like that in equation (7.1), there is no correlation between the error term and explanatory variables. That is:

$$Cov(u_i, X_{2i}) = Cov(u_i, X_{3i}) = 0 \quad \ldots (7.2)$$

In order to avoid this problem, we assume that all explanatory variables are non-stochastic in nature. This implies that the values taken by the explanatory variables X are considered fixed in repeated samples.

(iii) The mean of the error terms is zero. In other words, the expected value of the error term conditional upon the explanatory variables X_{2i} and X_{3i} is zero. This means:

$$E(u_i) = 0 \text{ or } E(u_i|X_{2i}, X_{3i}) = 0 \quad \ldots (7.3)$$

(iv) No autocorrelation: This assumption means that there is no serial correlation or autocorrelation between the error terms of the individual observations. This implies that the covariance between the error term associated with the i^{th} observation u_i and that with the j^{th} observation u_j is zero. In notations, this means:

$$\text{cov}(u_i, u_j) = 0 \quad \ldots (7.4)$$

(v) Homoscedasticity: The assumption of homoscedasticity implies that the error variance is constant for all observations. This means:

$$var(u_i^2) = \sigma^2 \quad \ldots (7.5)$$

(vi) No exact collinearity between the X variables. This is the new additional assumption made for multiple regression models. This implies that there is no exact linear relationship between X_2 and X_3 This is referred to as the assumption of no perfect multicollinearity.

(vii) The number of observations n must be greater than the number of parameters to be estimated. In other words, the number of observations n must be greater than the number of explanatory variables k.

(viii) No specification bias: It is assumed that the model is correctly specified. The assumption of no specification bias implies that there are no errors involved while specifying the model. This means that both the errors of including an irrelevant variable and not including a relevant variable are taken care of while specifying the regression model.

(ix) There is no measurement error, i.e., X's and Y are correctly measured.

7.2.1 Interpretation of the Model

In the multiple regression model as in equation (7.1), the intercept β_1 measures the expected value of the dependent variable Y, when the values of explanatory variables X_2 and X_3 are zero. The other two parameters, β_2 and β_3, are the partial regression coefficients. Let us know more about these coefficients. The regression coefficients β_2 and β_3 are also known as the partial slope coefficients. β_2 measures the change in the mean value of Y [i.e., E(Y)] per unit change in X_2, holding the value of X_3 constant. This means: $\beta_2 = \frac{\Delta E(Y)}{\Delta X_2}$. It gives the 'direct' or the 'net' effect of a unit change in X_2 on the mean value of Y holding the effect of X_3 constant. Likewise, β_3 measures the change in the mean value of Y, per unit change in X_3, holding the value of X_2 constant. Like β_2, β_3 is given by: $\beta_3 = \frac{\Delta E(Y)}{\Delta X_3}$. Thus, the slope coefficients of multiple regression measures the impact of

one explanatory variable on the dependent variable keeping the effect of the other variables fixed.

7.3 ESTIMATION OF MULTIPLE REGRESSION MODEL

The multiple regression equation is estimated to describe the Population Regression Function (PRF): $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$. This function consists of two components. The first is the deterministic component given by E ($Y_i \mid X_{2i}, X_{3i}$). This is also referred to as the Population Regression Line. The second component is the random component given by u_i. The PRF is estimated by using the sample. The estimated function (i.e., the sample regression function) is indicated by: $Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + e_i$. Recall that $Y_i = \hat{Y}_i + e_i$ where $\hat{Y}_i$ is the estimated value of Y_i given by $E(Y_i|X_{2i}, X_{3i})$ and e_i is the residual term. In the sample regression function, b_1 is the estimator of population intercept β_1 and b_2 and b_3 are the estimators of population partial slope coefficient β_2 and β_3 respectively. The residual e_i is the estimator of population error term u_i. We know that the sample regression line is obtained in the OLS method by minimizing the residual sum of squares as follows:

$$Min \sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2$$

$$= \sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \text{ [since } \hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i}]$$

We now consider the three first order conditions, i.e., $\frac{\partial \sum e_i^2}{\partial b_1} = 0$, $\frac{\partial \sum e_i^2}{\partial b_2} = 0$ and $\frac{\partial \sum e_i^2}{\partial b_3} = 0$. From these three partial derivatives, we obtain the estimators as:

(i) $b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3$

(ii) $b_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$

(iii) $b_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$

The corresponding variances and standard errors of the parameters are given as:

$$V(b_1) = \left[\frac{1}{n} + \frac{\bar{X}_2^2 \sum x_{3i}^2 + \bar{X}_3^2 \sum x_{2i}^2 - 2\bar{X}_2 \bar{X}_3 \sum x_{2i} x}{\sum x_{2i}^2 \sum x_{3i}^2 - (\sum x_{2i} x_{3i})^2}\right] \sigma^2$$

$$SE(b_1) = +\sqrt{V(b_1)}$$

$$V(b_2) = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \times \sigma^2$$

$$SE(b_2) = +\sqrt{V(b_2)}$$

$$V(b_3) = \frac{\sum x_{2i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})} \times \sigma^2$$

$$V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2(1-r_{23}^2)}$$

$$SE(b_3) = +\sqrt{V(b_3)}$$

You should note further that:

(i) $COV(b_2, b_3) = \frac{-r_{23}\sigma^2}{(1-r_{23}^2)\sqrt{x_{2i}^2}\sqrt{x_{3i}^2}}$

and the estimates of error variance and the partial correlation coefficients are given by:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} = \frac{RSS}{n-k} \quad \ldots (7.6)$$

For a regression model with 3 explanatory variables (such as equation (7.1)) we have $\hat{\sigma}^2 = \frac{RSS}{n-3}$.

$$r_{23} = \frac{(\sum x_{2i}x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2} \quad \ldots (7.7)$$

Note that in the above expressions, lower case letters represent deviations from the mean. We know that, since we are considering the 'classical' linear multiple regression model, the OLS estimators of the intercept and the partial slope coefficients satisfy the following properties:

a) The regression line passes through the means, $\bar{Y}, \bar{X}_2$ and $\bar{X}_3$. In a *k*-variable linear regression model, there is one regressand Y_i and $(k - 1)$ regressors since one of the coefficients is the intercept term β_1. Hence, the estimate of this intercept term is obtained as: $b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$.

b) The mean value of the estimated $\widehat{Y}\imath$ is equal to the mean value of actual Y_i, i.e., $\overline{\hat{Y}}_i = \overline{Y}$.

c) $\frac{1}{n}\sum e_i = \bar{e}_i = 0$.

d) $Cov(e_i, X_{2i}) = Cov(e_i, X_{3i}) = 0$. That is, the residual e_i is uncorrelated with X_{2i} and X_{3i}. In other words: $(\sum e_i X_{2i}) = (\sum e_i X_{3i}) = 0$.

e) $Cov\left(e_i, \hat{Y}_i\right) = 0$, i.e., residual e_i is uncorrelated with $\hat{Y}_i$ and $\sum e_i \hat{Y}_i = 0$.

f) As r_{23}, the correlation coefficient between X_2 and X_3, increases towards 1, the variances of b_2 and b_3 increases for given values of $\sigma^2, \sum x_{2i}^2$ or $\sum x_{3i}^2$.

g) In view of f) above, given the values of r_{23} and $\sum x_{2i}^2$ or $\sum x_{3i}^2$ the variances of OLS estimators are directly proportional to σ^2.

h) Given the assumptions of CLRM, OLS estimators of partial regression coefficients are not only linear and unbiased but also have minimum variances in the class of all unbiased estimators, i.e., they are BLUE. In other words, they satisfy the Gauss-Markov theorem.

7.4 MAXIMUM LIKELIHOOD METHOD OF ESTIMATION

The method of 'maximum likelihood estimation' estimates the parameters of a probability distribution function (pdf). This is done by maximizing the likelihood function of the pdf. Hence, the estimators that maximize the likelihood function are called the 'maximum likelihood estimators'. To understand this concept better, let us derive the maximum likelihood estimators ($\tilde{\beta}$). We have used the notation $\tilde{\beta}$ to distinguish the ML estimators from the OLS estimators ($\hat{\beta}$). Let us assume that the pdf follows normal distribution. Thus, $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. Taking log of the likelihood function of this pdf on its both sides, we get:

$$lnL = -\frac{n}{2}ln\sigma^2 - \frac{n}{2}ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i - \beta_1 - \beta_2 X_{2i} - \beta_k X_{ki})^2}{\sigma^2}$$

Differentiating the above function partially with respect to $\beta_1, \beta_2, ..., \beta_k$ and σ^2 we obtain the following (k + 1) equations:

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{1}{\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})\,(-1) \qquad (1)$$

$$\frac{\partial \ln L}{\partial \beta_2} = \frac{1}{\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})\,(-X_{2i}) \qquad (2)$$

……

$$\frac{\partial \ln L}{\partial \beta_k} = \frac{1}{\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})\,(-X_{ki}) \qquad (k)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_{2i} - \ldots - \beta_k X_{ki})^2 \qquad (k+1)$$

Setting these equations to zero (i.e., applying the first-order condition for optimization), and re-arranging terms, and denoting by $\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_k$ and $\tilde{\sigma}^2$ as the 'maximum likelihood estimates (MLEs)', we get:

$$\sum Y_i = n\tilde{\beta}_1 + \tilde{\beta}_2\sum X_{2i} + \cdots + \tilde{\beta}_k\sum X_{ki}$$

$$\sum Y_i X_{2i} = \tilde{\beta}_1 + \sum X_{2i} + \tilde{\beta}_2\sum X_{2i}^2 + \ldots + \tilde{\beta}_k\sum X_{2i}X_{ki}$$

……

$$\sum Y_i X_{ki} = \tilde{\beta}_1\sum X_{ki} + \tilde{\beta}_2\sum X_{2i}X_{ki} + \ldots + \tilde{\beta}_k\sum X_{ki}^2$$

The above equations are precisely the normal equations of the OLS method of estimation. Therefore, the MLEs of the $\tilde{\beta}'s$ are the same as the OLS estimates of the $\tilde{\beta}'s$. Thus, substituting the MLEs (or the OLS estimators) into the $(K+1)^{st}$ equation above, and simplifying, we obtain the MLEs of σ^2 as

$$\tilde{\sigma}^2 = \frac{1}{n}\sum\left(Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_{2i} - \ldots - \tilde{\beta}_k X_{ki}\right)^2$$

$$= \frac{1}{n}\sum \hat{u}_i^2$$

You may note that this estimator differs from the OLS estimator $\hat{\sigma}^2 = \sum u_i^2/(n-k)$. Since the latter is an unbiased estimator of σ^2, the MLE of $\tilde{\sigma}^2$ is a biased estimator. However, you should note that, asymptotically, $\tilde{\sigma}^2$ is also unbiased. This means, asymptotically, the estimates of MLE and OLS are similar. Further, the MLE estimator is biased but it is consistent.

For multiple regression models, the above algebraic expressions become unwieldy. Hence, we can take recourse to matrix algebra (on which you have studied in your earlier course BECC 104) to depict the multiple regression model. For this, let:

$$X_0 = \begin{bmatrix} 1 \\ X_{02} \\ X_{03} \\ \cdot \\ \cdot \\ \cdot \\ X_{0k} \end{bmatrix} \quad \ldots (7.8)$$

be the vector of values of the X variables for which we wish to predict $\hat{Y}_0$ the mean prediction of Y. Now the estimated multiple regression equation in the scalar form is:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \ldots + \hat{\beta}_k X_{ki} + u_i \quad \ldots (7.9)$$

In matrix notation (7.9) can be written compactly as:

$$\hat{Y}_i = x'_i \hat{\beta} \quad \ldots (7.10)$$

where $x'_i = [1 \quad X_{2i} \quad X_{3i} \ldots \quad X_{ki}]$ and

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_k \end{bmatrix}$$

Equation (7.9) or (7.10) is the mean prediction of Y_i corresponding to given x'_i. Hence, if x_i' is as given in (7.8), (7.10) becomes

$$\left(\hat{Y}_I|x'_0\right) = x'_0\hat{\beta} \quad \ldots (7.11)$$

where, the values of x_0 are specified. Note that (7. 11) gives an unbiased prediction of $E(Y_i|x'_0)$, since $E\left(x'_0\hat{\beta}\right) = x'_0\hat{\beta}$. The estimate of the variance of $\left(\hat{Y}_0|x'_0\right)$ is given by:

$$Var\left(\hat{Y}_0|x'_0\right) = \sigma^2 x'_0(X'X)^{-1}x_0 \quad \ldots (7.12)$$

where σ^2 is the variance of u_i, x'_0 are the given values of the X variables for which we wish to predict the future values, and $(X'X)$ is the matrix. In practice, we replace σ^2 by its unbiased estimator $\hat{\sigma}^2$.

Check Your Progress 1 [answer the questions in 50-100 words within the given space]

1) Specify the simplest form of a multiple regression model with examples. Why is it the simplest?

.....
.....
.....
.....
.....

2) Enumerate the assumptions made for the CLRM in broad terms. What is the additional assumption made for the multiple regression model?

.....
.....
.....
.....
.....

3) How are the estimated parameters of a multiple regression model interpreted?

.....
.....
.....
.....
.....

4) Specify the satisfaction of the property which makes the OLS estimators obey the Gauss Markov theorem?

.....
.....
.....
.....
.....

7.5 COEFFICIENT OF DETERMINATION: R^2

In multiple regression, a measure of goodness of fit is given by R^2. This is also called as the 'coefficient of determination'. It is the ratio of the 'explained sum of squares' to the 'total sum of squares'. In other words, it is the proportion of total variation in the dependent variable explained by the independent (or the explanatory) variables included in the model. To derive R^2, we consider the sample regression function or equation as follows:

$$Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + e_i \quad \ldots (7.13)$$

where $b_1 = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3$. Substituting b_1 in (7.13), and by considering X_{2i} and X_{3i} in their means, we get:

$$Y_i = \bar{Y} - b_2\bar{X}_2 - b_3\bar{X}_3 + b_2X_{2i} + b_3X_{3i} + e_i$$

Therefore, $Y_i - \bar{Y} = b_2(X_{2i} - \bar{X}_2) + b_3(X_{3i} - \bar{X}_3) + e_i$

Rewriting the above in lower case, i.e., by considering in deviation from mean, we get:

$$y_i = b_2x_{2i} + b_3x_{3i} + e_i \quad \ldots (7.14)$$

We have $\hat{Y}_i - \overline{Y} = \hat{Y}_i$ where:

$$\hat{Y}_i = b_1 + b_2X_{2i} + b_3X_{3i}$$

$$\overline{Y} = b_1 + b_2\overline{X}_2 + b_3\overline{X}_3$$

$$\therefore \ \hat{Y}_i - \overline{Y} = \left(b_1 + b_2X_{2i} + b_3X_{3i}\right) - \left(b_1 + b_2\overline{X}_2 + b_3\overline{X}_3\right)$$

$$\hat{Y}_i - \overline{Y} = b_2\left(X_{2i} - \overline{X}_2\right) + b_3\left(X_{3i} - \overline{X}_3\right)$$

$$\hat{Y}_i - \overline{Y} = b_2x_{2i} + b_3x_{3i} \quad \ldots\ldots (7.15)$$

Now, consider:

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

Squaring both sides and summing up we get

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i$$

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 0 \ \text{ since } \ Cov\left(\hat{y}_i, e_0\right) = 0$$

$$\therefore \sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \qquad \ldots (7.16)$$

It means TSS = ESS + RSS. Now, consider: $R^2 = \frac{ESS}{TSS}$ where ESS = $\sum \hat{y}_i^2$.

Since $e_i = y_i - \hat{y}_i$ with $\hat{y}_i = b_2 x_{2i} + b_3 x_{3i}$ we have: $e_i = y_i - (b_2 x_{2i} + b_3 x_{3i})$

Now, $\sum e_i^2 = \sum (e_i e_i)$

$= \sum [e_i (y_i - b_2 x_{2i} - b_3 x_{3i})]$

$= \sum e_i y_i - b_2 \sum e_i x_{2i} - b_3 \sum e_i x_{3i}$

$\therefore \sum e_i^2 = \sum e_i y_i$ [since $\sum e_i x_{2i} = \sum e_i x_{3i} = 0$].

$\sum e_i^2 = \sum y_i e_i = \sum y_i (y_i - b_2 x_{2i} - b_3 x_{3i})$

$\Rightarrow \sum e_i^2 = \sum y_i^2 - b_2 \sum y_i x_{2i} - b_3 \sum y_i x_{3i} \qquad \ldots (7.17)$

Using (7.17) in (7.16) we get:

$\sum y_i^2 = \sum \hat{y}_i^2 + \sum y_i^2 - b_2 \sum y_i x_{2i} - b_3 \sum y_i x_{3i}$

$\Rightarrow \sum \hat{y}_i^2 = b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}$ = ESS

Therefore, $R^2 = \frac{ESS}{TSS} = \frac{b_2 \sum y_i x_{2i} + b_3 \sum y_i x_{3i}}{\sum y_i^2} \qquad \ldots (7.18)$

The relationship between R^2 and variance of a partial regression coefficient (b_i) in a *k*-variable multiple regression model is given by:

$$V(b_i) = \frac{\sigma^2}{\sum x_y^2} - \left(\frac{1}{1 - R_i^2}\right)$$

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

$$= 1 - \frac{(n-k)\hat{\sigma}^2}{(n-1)Sy^2}$$

$$\because \sum e_i^2 \text{ or } \hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} \Rightarrow \sum e_i^2 = (n-k)\hat{\sigma}^2$$

$$Sy^2 = \frac{\sum y_i^2}{n-1} \Rightarrow \sum y_i^2 = (n-1)Sy^2$$

7.6 ADJUSTED-R^2

In comparing two regression models with the same dependent variable but differing number of X variables, one should be careful in choosing the model with highest R^2. In order to understand why this is important, consider:

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

Note that as the number of explanatory variables increase, the numerator ESS keeps on increasing. In other words, R^2 increases as k, the number of independent variables increase. The above expression for R^2 implies that R^2 does not give any weightage to the number of independent variables in the model. Due to this reason, for comparison of two regressions with differing number of explanatory variables, we should not use R^2. We now need an alternative coefficient of determination which takes into account the number of parameters estimated, i.e., k. For this, we consider the following measure called the adjusted R^2 defined as follows.

$$\bar{R}^2 = 1 - \frac{RSS/n-k}{TSS/n-1}$$

$$= 1 - \frac{\sum e_i^2/(n-k)}{\sum y_i^2/(n-1)}$$

where k is the number of parameters in the model including the intercept term. The above is same as saying:

$$\bar{R}^2 = \frac{1-\hat{\sigma}^2}{S_y^2}$$

where $\hat{\sigma}^2$ is the residual variance which is an unbiased estimator of true σ^2. S_y^2 is the sample variance of Y. Now, a relationship between $\bar{R}^2$ and R^2 is given by

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \quad \ldots (7.19)$$

Now, for deciding on whether R^2 or $\bar{R}^2$ should be used, we must note the following:

(i) If $k > 1, \bar{R}^2 < R^2$. This implies that as the no. of explanatory variables X increases, the adjusted R^2 increases less than the usual R^2

(ii) $\bar{R}^2$ can be negative but R^2 is necessarily non-negative. This is because, in (7.18):

If $R^2 = 1, \bar{R}^2 = 1$.

If $R^2 = 0$, $\bar{R}^2 = \frac{1-k}{n-k}$. Hence, if $k > 1$ then $\bar{R}^2 < 0$.

Thus, adjusted R^2 can be negative. In such cases, it is conventional to take the value of $\overline{R}^2$ as zero. Thus, a conclusive opinion on which of the two is superior to indicate the goodness of fit of a regression model is not possible. However, in practice, in multiple regression models, adjusted R^2 is used to decide for the goodness of fit of the model for the reason that it takes into account the number of regressors and thereby the number of parameters estimated.

Check Your Progress 2 [answer the questions in 50-100 words within the given space]

1) Distinguish between the OLS estimate and the MLE.

.....
.....
.....
.....
.....

2) How is R^2 defined? Indicate with suitable expressions.

.....
.....
.....
.....
.....
.....

3) State the importance of adjusted-R^2 as compared to R^2.

.....
.....
.....
.....
.....

4) How are R^2 and adjusted-R^2 related? What is the difference between the two?

.....
.....
.....
.....
.....

5) How is the situation of adjusted-R^2 being negative dealt with in practice?

.....
.....
.....
.....
.....

7.7 LET US SUM UP

This unit has described the multiple regression model and its inferences. Recapitulating the assumptions of the multiple classical regression model, the unit indicates how an additional assumption on multicollinearity is necessary in multiple regression models. The interpretation of parameters, i.e., the intercept and the partial slope coefficient are explained. The unit has first discussed the estimation of parameters of the multiple regression model by the OLS (ordinary least squares) method. An alternative method, namely the method of maximum likelihood estimation (MLE) is introduced in the unit next. It is shown that asymptotically the OLS and the MLE coincide. The concept of 'coefficient of determination' or goodness of fit has been described. Finally, the need and the use of adjusted R^2 has been explained.

7.8 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) A multiple regression model is one in which there is more than one independent or the explanatory variable. Hence, the simplest multiple regression model is one with one dependent variable and two independent variables. Such a model is specified as: $\boldsymbol{Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i}$. Examples can be a production function in which the dependent variable is the output and the independent variables are two inputs viz. labour and capital. In microeconomics, it could be a relationship between consumption of a commodity as the dependent variable and price and income as the two independent variables.

2) (i) The model is linear in parameters; (ii) u_i and X_i are not correlated, i.e., $\boldsymbol{cov(u_i, X_{2i}) = Cov(u_i, X_{3i}) = 0}$; (iii) the conditional expectation of the error term is zero, i.e., $\boldsymbol{E(u_i | X_{2i}, X_{3i}) = 0}$; (iv) error terms are not correlated or there is no auto correlation, i.e., $\boldsymbol{\mathbf{cov}(u_i, u_j) = 0}$; (v) there is homoscedasticity or the error variance do not differ, i.e., $\boldsymbol{var(u_i^2) = \sigma^2}$; (vi) no multicollinearity or perfect collinearity, i.e., *Corr* (X_i, X_j) ≠ 1; (vii) number of observations (n) is greater than the number of parameters estimated (k); (viii) there is no specification bias, i.e., neither a relevant variable is omitted nor an irrelevant variable is included in the model; and (ix) there is no measurement error in X's and Y. Assumption no (vi) above is the additional assumption required in multiple regression models.

3) The intercept β_1 measures the expected value of the dependent variable Y, given the values of explanatory variables X_2 and X_3. β_2 measures the change in the mean value of Y [i.e., E(Y)] per unit change in X_2, holding the value of X_3 constant. This means: $\beta_2 = \frac{\Delta E(Y)}{\Delta X_2}$. Likewise, β_3 is defined.

4) Under the assumptions of CLRM, the OLS estimators of partial regression coefficients are not only linear and unbiased but also have minimum variances in the class of all unbiased estimators, i.e., they are BLUE (best liner unbiased estimate). It is this property that makes the OLS estimates satisfy the Gauss-Markov theorem.

1) Check Your Progress 2The OLS estimators are obtained by minimizing the residual sum of squares, i.e., Min $\sum e_i^2 = \sum\left(Y_i - \hat{Y}_i\right)^2$. The MLEs are obtained by maximising the 'likelihood function' of the corresponding pdf. There is thus a basic difference in the approach of the two methods. However, once the first order conditions are applied and simplified, the equations that we obtain in the MLE approach is same as the normal equations that we get in the OLS method. Hence, the estimates for the parameters obtained by solving those equations are the same. However, there is an essential difference relating to the unbiased estimate of σ^2. The denominator of the expression for this unbiased estimate in the OLS method is '*n-k*' whereas in the ML method it is '*n*'. This important difference makes the estimate of σ^2 in the ML approach biased for small samples. For large samples, it is unbiased. Hence, the estimates of ML and OLS are similar and asymptotically, the OLS and the MLEs coincide.

2) For a 2 independent variables multiple regression model, whose sample regression function is given as $Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + e_i$ the R^2 is defined as: $R^2 = \dfrac{ESS}{TSS} = \dfrac{b_2\sum y_i x_{2i} + b_3\sum y_i x_{3i}}{\sum y_i^2}$.

(iii) For comparing two multiple regressions with differing number of explanatory variables, relying on R^2 could be misleading. This is because R^2 does not take into account the number of explanatory variables.

(iv) They are related as: $\overline{R}^2 = 1 - \left(1 - R^2\right)\dfrac{n-1}{n-k}$. An important difference is that while R^2 cannot be negative, adjusted R^2 can be negative.

5) When this is negative, conventionally it is taken as zero.

UNIT 8 MULTIPLE LINEAR REGRESSION MODEL: INFERENCES*

Structure

8.0 OBJECTIVES

After going through this unit, you should be able to

- explain the need for the assumption of normality in the case of multiple regression;
- describe the procedure of testing of hypothesis on individual estimators;
- test the overall significance of a regression model;
- test for the equality of two regression coefficients;
- explain the procedure of applying the Chow test;
- make prediction on the basis of multiple regression model;
- interpret the results obtained from the testing of hypothesis, both individual and joint; and
- apply various tests such as likelihood ratio (LR), Wald (W) and Lagrange Multiplier Test (LM).

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

8.1 INTRODUCTION

In the previous unit we discussed about the interpretation and estimation of multiple regression models. We looked at the assumptions that are required for the ordinary least squares (OLS) and maximum likelihood (ML) estimation. In the present Unit we look at the methods of hypothesis testing in multiple regression models.

Recall that in Unit 3 of this course we mentioned the procedure of hypothesis testing. Further, in Unit 5 we explained the procedure of hypothesis testing in the case of two variable regression models. Now let us extend the procedure of hypothesis testing to multiple regression models. There could be two scenarios in multiple regression models so far as hypothesis testing is concerned: (i) testing of individual coefficients, and (ii) joint testing of some of the parameters. We discuss the method of testing for structural stability of regression model by applying the Chow test. Further, we discuss three important tests, viz., Likelihood Ratio test, Wald test, and Lagrange Multiplier test. Finally, we deal with the issue of prediction on the basis of multiple regression equation.

One of the assumptions in hypothesis testing is that the error variable u_i follows normal distribution. Is there a method to test for the normality of a variable? We will discuss this issue also. However, let us begin with an overview of the basic assumptions of multiple regression models.

8.2 ASSUMPTIONS OF MULTIPLE REGRESSION MODELS

In Unit 7 we considered the multiple regression model with two explanatory variables X_2 and X_3. The stochastic error term is u_i.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \ldots (8.1)$$

8.2.1 Classical Assumptions

There are seven assumptions regarding the multiple regression model. Most of these assumptions are regarding the error term. We discussed about these assumptions in the previous Unit. Let us briefly mention those assumptions again.

a) The regression model is linear in parameters and variables.

b) The mean of error terms is zero. In other words, the expected value of error term conditional upon the explanatory variables X_{2i} and X_{3i} is zero.

$$E(u_i) = 0 \text{ or } E(u_i | X_{2i}, X_{3i}) = 0$$

c) There is no serial correlation (or autocorrelation) among the error terms. The error terms are not correlated. It implies that the covariance between the error term associated with i^{th} observation u_i and the error term associated with j^{th} observation, u_j is zero.

$$\text{cov}(u_i, u_j) = 0$$

d) Homoscedasticity: The assumption of homoscedasticity states that the error variance is constant throughout the population. The variance of the error term associated at each observation has the same variance.

$$var(u_i) = \sigma^2$$

e) Exogeneity of explanatory variables: There is no correlation between the explanatory variables and the error term. This assumption is also called exogeneity, because the explanatory variables are assumed to be exogenous (given from outside; X is not determined inside the model). In contrast, Y is determined within the model. When the explanatory variable is correlated with the error term, it is called endogeneity problem. In order to avoid this problem, we assume that the explanatory variables are kept fixed across samples.

f) Independent variables are not linear combination of one another. If there is perfect linear relationship among the independent variables, the explanatory variables move in harmony and it is not possible to estimate the parameters. It is also called multicollinearity problem.

g) The error variable is normally distributed. This assumption is not necessary in OLS method for estimation of parameters. It is required for construction of confidence interval and hypothesis testing. In the maximum likelihood method discussed in the previous Unit, in order to estimate the parameters we assumed that the error term follows normal distribution.

8.2.2 Test for Normality of the Error Term

As pointed out earlier, we look into the assumption of normality of the error term. In order to test for normality of the error term we apply the Jarque-Bera test (often called the JB test). It is an asymptotic or large sample test. We do not know the error terms in a regression model; we know the residuals. Therefore, the JB test is based on the OLS residuals. Recall two concepts from statistics: skewness and kurtosis. A skewed curve (i.e., asymmetric) is different from a normal curve. A leptokurtic or platykurtic curve (i.e., tall or short in height) is different from a normal curve. The JB test utilises the measures of skewness and kurtosis.

We know that for a normal distribution S = 0 and K = 3. A signification deviation from these two values will confirm that the variable is not normally distributed.

Jarque and Bera constructed the J-statistic given by

$$JB = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right] \quad \ldots (8.2)$$

where

n = sample size

S = measure of skewness ($\frac{\mu_3}{\sigma^3}$)

K = measure of kurtosis ($\frac{\mu_4}{\mu_2^2}$)

Skewness and kurtosis are measured in terms of the moments of a variable. As you know from BECC 107, Unit 4, the formula for calculating the rth moment of variable X_i is

$$\mu_r = \frac{1}{n}\sum_{i-1}^{n} f_i(X_i - \bar{X})^r \quad \ldots (8.4)$$

Variance is the second moment μ_2.

In equation (8.2) the JB statistic follows chi-square distribution with 2 degrees of freedom, $\sim\chi^2_{(2)}$.

Let us find out the value of the JB statistic if a variable follows normal distribution. For the normal distribution, as mentioned above S = 0 and K = 3. By asubstitutingthese values in equation (8.2) we obtain

$$JB = \frac{n}{6}[0 + 0] = \frac{n}{6} \times 0 = 0 \quad \ldots (8.3)$$

For a variable not normally distributed JB statistics will assume increasingly large values. The null hypothesis is

H_0: The random variable follows normal distribution.

We draw inferences from the JB statistic as follows:

a) If the calculated value of JB statistic is greater than the tabulated value of χ^2 for 2 degrees of freedom, we reject the null hypothesis. We infer that the random variable is not normally distributed.

b) If the calculated value of the JB statistic is less than the tabulated value of χ^2 for 2 degrees of freedom, we do not reject the null hypothesis. We infer that the random variable is normally distributed.

Check Your Progress 1

1) List the assumptions of multiple regression models.

.....

.....

.....

.....

2) State the Jarque-Bera test for normality.

.....

.....

.....

.....

8.3 TESTING OF SINGLE PARAMETER

The population regression function is not known to us. We estimate the parameters on the basis of sample data. Since we do not know the error variance σ^2, we should apply *t*-test instead of *z*-test (based on normal distribution).

Let us consider the population regression line given at equation (8.1).

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

The sample regression line estimated by ordinary least squares (OLS) method is

$$\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i} \quad \ldots (8.4)$$

where b_1, b_2 and b_3 are estimators of β_1, β_2 and β_3 respectively. The estimator of error variance σ^2 is given by $\hat{\sigma}^2 = \frac{RSS}{n-k}$.

There are two approaches to hypothesis testing: (i) test of significance approach, and (ii) confidence interval approach. We discuss both the approaches below.

8.3.1 Test of Significance Approach

In this approach we proceed as follows:

(i) Take the point estimate of the parameter that we want test, viz., b_1, or b_2 or b_3.

(ii) Set the null hypothesis. Suppose we expect that variable X_2 has no influence on *Y*. It implies that β_2 should be zero. Thus, null hypothesis is $H_0: \beta_2 = 0$. In this case what should be alternative hypothesis? The alternative hypothesis is $H_A: \beta_2 \neq 0$.

(iii) If $\beta_2 \neq 0$, then $\beta_2$could be either positive or negative. Thus we have to apply two-tail test. Accordingly, the critical value of the t-ratio has to be decided.

(iv) Let us consider another scenario. Suppose we expect that $\beta_3$should be positive. It implies that our null hypothesis is $H_0: \beta_3 > 0$. The alternative hypothesis is $H_A: \beta_3 \leq 0$.

(v) If $\beta_3 > 0$, then β_3 could be either zero or negative. Thus the critical region or rejection region lies on one side of the *t* probability curve. Therefore, we have to apply one-tail test. Accordingly the critical value of t-ratio is to be decided.

(vi) Remember that the null hypothesis depends on economic theory or logic. Therefore, you have to set the null hypothesis according to some logic. If you expect that the explanatory variable should have no effect on the dependent variable, then set the parameter as zero in the null hypothesis.

(vii) Decide on the level of significance. It represents extent of error you want to tolerate. If the level of significance is 5 per cent (α = 0.05),

your decision on the null hypothesis will go be wrong 5 per cent times. If you take 1 per cent level of significance (α = 0.01), then your decision on the null hypothesis will be wrong 1 per cent times (i.e., it will be correct 99 per cent times).

(viii) Compute the t-ratio. Here the standard error is the positive square root of the variance of the estimator. The formula for the variance of the OLS estimators in multiple regression models is given in Unit 7.

$$t = \frac{b_2 - \beta_2}{se(b_2)} \quad \ldots (8.5)$$

(ix) Compare the computed value of the t-ratio with the tabulated value of the t-ratio. Be careful about the two issues while reading the t-table: (i) level of significance, and (ii) degree of freedom. Level of significance we have mentioned above. Degree of freedom is (n–k), as you know from the previous Unit.

(x) If the computed value of t-ratio is greater than the tabulated value of t-ratio, reject the null hypothesis. If computed value of t-ratio is less than the tabulated value of t-ratio, do not reject the null hypothesis and accept the alternative null hypothesis.

8.3.2 Confidence Interval Approach

We have discussed about interval estimation in Unit 3 and Unit 5. Thus, here we bring out the essential points only.

(i) Remember that confidence interval (CI) is created individually for each parameter. There cannot be a single confidence interval for a group of parameters.

(ii) Confidence interval is build on the basis of the logic described above in the test of significance approach.

(iii) Suppose we have the null hypothesis $H_0: \beta_2 = 0$ and the alternative hypothesis is $H_A: \beta_2 \neq 0$. The estimator of β_2 is b_2. We know the standard error of b_2.

(iv) Here also we decide on the level of significance (α). We refer to the t-table and find out the t-ratio for desired level of significance.

(v) The degree of freedom is known to us, i.e., (n–k).

(vi) Since the above is case of two-tailed test, we take $\alpha/2$ on each side of the t probability curve. Therefore, we take the t-ratio corresponding to the probability $\alpha/2$ and the degrees of freedom applicable.

(vii) Remember that confidence interval is created with the help of the estimator and its standard error. We test whether the parameter lies within the confidence interval or not.

(viii) Construct the confidence interval as follows:

$$\left[b_2 - t_{\alpha/2}SE(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2}SE(b_2)\right]. \quad \ldots (8.6)$$

(ix) The probability of the parameter remaining in the confidence interval is $(1-\alpha)$. If we have taken the confidence interval as 5 per cent, then the probability that β_2 will remain in the confidence interval is 95 per cent.

$$P_r\left[b_2 - t_{\alpha/2}SE(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2}SE(b_2)\right] = (1-\alpha) \ldots (8.7)$$

(x) If the parameter (in this case, β_2) remains in the confidence interval, do not reject the null hypothesis.

(xi) If the parameter does not remain within the confidence interval, reject the null hypothesis, and accept the alternative null hypothesis.

Check Your Progress 2

1) Describe the steps you would follow in testing the hypothesis that $\beta_2 < 0$.

.....

.....

.....

.....

.....

2) Create a confidence interval for the population parameter of the partial slope coefficient.

.....

.....

.....

.....

.....

8.4 TEST OF OVERALL SIGNIFICANCE

The overall test of significance of a multiple regression model is carried out by applying *F*-test. We have discussed about the *F*-test in Unit 5 of this course in the context of two variable models. For testing of the overall significance of a multiple regression model we proceed as follows:

(i) Set the null hypothesis. The null hypothesis for testing the overall significance of a multiple regression model is given as follows:

$$H_0: \beta_2 = \beta_3 = \ldots \beta_k = 0 \quad \ldots (8.8)$$

(ii) Set the corresponding alternative hypothesis.

$$H_A: \beta_2 = \ldots = \beta_k \neq 0 \quad \ldots (8.9)$$

(iii) Decide on the level of significance. It has the same connotation as in the case of *t*-test described above.

(iv) For multiple regression model the *F*-statistic is given by

$$F = \frac{ESS/(k-1)}{RSS(n-k)} \quad \ldots (8.10)$$

(v) Find out the degrees of freedom. The *F*-statistic mentioned in equation (8.10) follows *F* distribution with degrees of freedom (k–1, n–k).

(vi) Find out the computed value of *F* on the basis of equation (8.10). Compare it with the tabulated value of *F* (given at the end of the book). Read the tabulated *F* value for desired level of significance and applicable degrees of freedom.

(vii) If the computed value of *F* is greater than the tabulated value, then reject the null hypothesis.

(viii) If the computed value is less than the tabulated value, do not reject the null hypothesis.

8.5 TESTOF EQUALITY BETWEEN TWO PARAMETERS

We can compare between the parameters of a multiple regression model. Particularly, we can test whether two parameters are equal in a regression model. For this purpose we apply the same procedure as we have learnt in the course BECC 107.

Let us take the following regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i \quad \ldots (8.11)$$

Recall that we do not know the variance of the parameters. Thus, for comparison of the parameters we apply the *t*-test. Secondly, we do not the parameters. Therefore, we take their OLS estimators for comparison purposes.

Our null hypothesis and alternative hypothesis are as follows:

$$H_0: \beta_3 = \beta_4 \quad \text{or} \quad (\beta_3 - \beta_4) = 0 \quad \ldots (8.12)$$

$$H_1: \beta_3 \neq \beta_4 \quad \text{or} \quad (\beta_3 - \beta_4) \neq 0 \quad \ldots (8.13)$$

For testing of the above hypothesis, the t -statistic is given as follows:

$$t = \frac{(b_3 - b_4) - (\beta_3 - \beta_4)}{SE(b_3 - b_4)} \quad \ldots (8.14)$$

The above follows *t*-distribution with (n – k) degrees of freedom.

Since $\beta_3 = \beta_4$ under the null hypothesis, we can re-arrange equation (8.14) as follows:

$$t = \frac{b_3 - b_4}{\sqrt{V(b_3) + V(b_4) - 2\text{cov}(b_3, b_4)}} \quad \ldots (8.15)$$

The computed value of t-statistic is obtained by equation (8.15). We compare the computed value of *t*-ratio with the tabulated value of *t*-ratio. We read the *t*-table for desired level of significance and applicable degrees of freedom.

If the computed value of *t*-ratio is greater than the tabulated value, then we reject the null hypothesis. If the computed value of *t*-ratio is less than the tabulated value, then we do not reject the null hypothesis and accept the alternative hypothesis.

We need to interpret our results. If we reject the null hypothesis we conclude that the partial slope coefficients β_3 and β_4 are statistically significantly different. If we do not reject the null hypothesis, we conclude that there is no statistically significant difference between the slope coefficients β_3 and β_4.

Check Your Progress 3

1) Mention the steps of carrying out a test of the overall significance a multiple regression model.

.....

.....

.....

.....

.....

2) State how the equality between two parameters can be tested.

.....

.....

.....

.....

.....

8.6 TEST OF LINEAR RESTRICTIONS ON PARAMETERS

Many times we come across situations where we have to test for linear restrictions on parameters. For example, let us consider the Cobb-Douglas production function.

$$Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i} \quad \ldots (8.16)$$

where Y_i is output, X_{2i} is capital and X_{3i} is labour. The parameters are β_2 and β_3. The stochastic error term is u_i. The subscript 'i' indicates the i^{th} observation. The Cobb-Douglas production function exhibits constant returns to scale if the parameters fulfil the following condition:

$$\beta_2 + \beta_3 = 1 \quad \ldots (8.17)$$

As we have discussed in Unit 6, by taking natural log, the Cobb-Douglas production function can be expressed in linear form as

$$ln\, Y_i = ln\, \beta_1 + \beta_2 lnX_{2i} + \beta_3 lnX_{3i} + u_i \quad \ldots (8.18)$$

Suppose have collected data on a sample of firms; our sample size is *n*. The production function is Cobb-Douglas as given above. We want to test whether the production function exhibits constant returns to scale. For this purpose we need to apply the F-test. We can follow two approaches as discussed below.

8.6.1 The t-Test Approach

We will discuss two procedures for testing the hypothesis.

(a) For the In this case our null hypothesis and alternative hypothesis are as follows:

$$H_0: \beta_2 + \beta_3 = 1 \quad \ldots (8.19)$$

$$H_A: \beta_2 + \beta_3 \neq 1 \quad \ldots (8.20)$$

For testing of the above hypothesis, the *t* -statistic is given as follows:

$$t = \frac{(b_2+b_3)-(\beta_2+\beta_3)}{SE(b_2+b_3)} \quad \ldots (8.21)$$

The above follows *t*-distribution with (n – k) degrees of freedom.

We can re-arrange equation (8.21) as follows:

$$\mathrm{t} = \frac{b_2+b_3-1}{\sqrt{V(b_2)+V(b_3)+2\mathrm{cov}(b_2,b_3)}} \quad \ldots (8.22)$$

The computed value of t-statistic is obtained by equation (8.22). We compare the computed value of *t*-ratio with the tabulated value of *t*-ratio. We read the *t*-table for desired level of significance and applicable degrees of freedom.

If the computed value of *t*-ratio is greater than the tabulated value, then we reject the null hypothesis. If the computed value of *t*-ratio is less than the tabulated value, then we do not reject the null hypothesis and accept the alternative hypothesis.

We need to interpret our results. If we reject the null hypothesis we conclude that the firms do not exhibit constant returns to scale. If we do not reject the null hypothesis, we conclude that the firms exhibit constant returns to scale.

(b) Let us look again at the null hypothesis given at (8.19).

$$H_0: \beta_2 + \beta_3 = 1$$

If the above restriction holds, then we should have

$$\beta_2 = (1 - \beta_3)$$

Let us substitute the above relationship in the Cobb-Douglas production function

$$ln\, Y_i = ln\, \beta_1 + (1 - \beta_3) lnX_{2i} + \beta_3 lnX_{3i} + u_i \quad \ldots (8.23)$$

We can re-arrange terms in equation (8.23) to obtain

$$ln\, Y_i - lnX_{2i} = ln\, \beta_1 - \beta_3 lnX_{2i} + \beta_3 lnX_{3i} + u_i$$

Or,

$$ln(Y_i/X_{2i}) = \beta_0 + \beta_3 ln(X_{3i}/X_{2i}) + u_i \quad \ldots (8.24)$$

Note that the dependent variable in the above regression model is output-labour ratio and the explanatory variable is capital-labour ratio. We can estimate the regression model given at equation (8.24) and find the OLS estimator of β_3.

If $\beta_3 = 1$, then the Cobb-Douglas production will exhibit constant returns to scale.

Therefore, we set the null hypothesis and alternative hypothesis as

$H_0: \beta_3 = 1$ and $H_A: \beta_3 \neq 1$

We apply t-test for individual parameters as mentioned in sub-section 8.3.1. If the null hypothesis is rejected we conclude that the firms do not exhibit constant returns to scale.

8.6.2 Restricted Least Squares

The t-test approach mentioned above may not be suitable in all cases. There may be situations where we have more than two parameters to be tested. In such circumstances we apply the *F*-test. This approach is called the restricted least squares.

Let us consider the multiple regression model given at equation (8.11).

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

Suppose we have to test the hypothesis that X_3 and X_4 do not influence the dependent variable Y. In such a case, the parameters β_3 and β_4 should be zero.

Recall that if we increase the number of explanatory variables in a regression model, there is an increase R^2. Recall further that $R^2 = \frac{ESS}{TSS}$. Thus, if two of the explanatory variables in equation (8.11) are dropped (i.e., their coefficients are zero), there will be a decrease in the value R^2. If the variables that X_3 and X_4 are relevant, there will be a significant decline the value of R^2. On the other hand, if the variables X_3 and X_4 are not relevant for the regression model, then the decline in the value of R^2 will be insignificant. We use this property of the regression model to test hypotheses on a group of parameters. Therefore, while applying F-test in restricted least squares we estimated the regression model twice: (i) the unrestricted model, and (ii) the restricted model.

We proceed as follows:

(i) Suppose there are *k* explanatory variables in the regression model.

(ii) Out of these *k* explanatory variables, suppose the first *m* explanatory variables are not relevant.

(iii) Thus our null hypothesis will be as follows:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0 \quad \ldots (8.25)$$

(iv) The corresponding alternative hypothesis will be that the βs are not zero.

(v) Estimate the unrestricted regression model given at (8.11). Obtain the residual sum of squares (RRS) on the basis of the estimated regression equation. Denote it as RSS_{UR}.

(vi) Estimate the restricted regression model by excluding the explanatory variables for which the parameters are zero. Obtain the residual sum of squares (RRS) from this restricted model. Denote it as RSS_R.

(vii) Our F-statistic is

$$F = \frac{RSS_R - RSS_{UR}/m}{RS_{UR}/(n-k)} \quad \ldots (8.26)$$

The *F*-statistic at (8.26) follows *F*-distribution with degrees of freedom (*m*, *n–k*).

(ix) Find out the computed value of *F* on the basis of equation (8.10). Compare it with the tabulated value of *F* (given at the end of the book). Read the tabulated *F* value for desired level of significance and applicable degrees of freedom.

(x) If the computed value of *F* is greater than the tabulated value, then reject the null hypothesis.

(xi) If the computed value is less than the tabulated value, do not reject the null hypothesis.

As mentioned earlier, the residual sum of squares (RSS) and the coefficient of determination (R^2) are related. Therefore, it is possible to carry out the *F*-test on the basis of R^2 also. If we have the coefficient of determination for the unrestricted model (R^2_{UR}) and the coefficient of determination for the restricted model (R^2_R), then we can test the joint hypothesis about the set of parameters.

The *F*-statistic will be

$$F = \frac{R^2_{UR} - R^2_R/m}{(1-R^2_{UR})/(n-k)} \quad \ldots (8.27)$$

which follows *F*-distribution with degrees of freedom (*m*, *n–k*).

The conclusion to be drawn and interpretation of results will be the same as described in points (x) and (xi) above.

8.7 STRUCTURAL STABILITY OF A MODEL: CHOW TEST

Many times we come across situations where there is a change in the pattern of data. The dependent and independent variables may not remain the same throughout the sample. For example, saving behaviour of poor and rich households may be different. The production of an industry may be different after a policy change. In such situations it may not be appropriate to run a single regression for the entire dataset. There is a need to check for structural stability of the econometric model.

There are various procedures to bring in structural breaks in a regression model. We will discuss about the dummy variable cases in unit 9. In this Unit we discuss a very simple and specific case.

Suppose we have data on n observations. We suspect that the first n_1 observations are different from the remaining n_2 observations (we have $n_1 + n_2 = n$). In this case run the following three regression equations:

$Y_t = \lambda_1 + \lambda_2 X_t + u_t$ (number of observations: n_1) … (8.28)

$Y_t = r_1 + r_2 X_t + v_t$ (number of observations: n_2) … (8.29)

$Y_t = \alpha_1 + \alpha_2 X_t + w_t$ (number of observations: $n = n_1 + n_2$) … (8.30)

If both the sub-samples are the same, then we should have $\lambda_1 = r_1 = \alpha_1$

and $\lambda_2 = r_2 = \alpha_2$. If both the sub-samples are different then there will be a structural break in the sample. It implies the parameters of equations (8.28) and (8.29) are different. In order to test for the structural stability of the regression model we apply Chow test.

We process as follows:

(i) Run the regression model (8.28). Obtain residual sum of squares RSS_1.

(ii) Run regression model (8.29). Obtain residual sum of squares RSS_2.

(iii) Run regression model (8.30). Obtain residual sum of squares RSS_3.

(iv) In regression model (8.30) we are forcing the model to have the same parameters in both the sub-samples. Therefore, let us call the residual sum of squares obtained from this model RSS_R.

(v) Since regression models given at (8.28) and (8.29) are independent, let us call this the unrestricted model. Therefore, $RSS_{UR} = RSS_1 + RSS_2$

(vi) Suppose both the sub-samples are the same. In that case there should not be any difference between RSS_{UR} and RSS_R. Our null hypothesis in that case is H_0: There is not structural change (or, there is parameter stability).

(vii) Test the above by the following test statistic:

$$F = \frac{RSS_R - RSS_{UR})/K}{RSS_{UR}/n_1+n_2-2k} \quad \ldots (8.31)$$

It follows F-distribution with degrees of freedom $k, (n_1 + n_2 - 2k)$, where k is the number of explanatory variables in the regression model.

(viii) Check the F-distribution table given at the end of the book for desired level of significance and applicable degrees of freedom.

(ix) Draw the inference on the basis of computed value of the F-statistic obtained at step(vii).

(x) If the computed value of *F* is greater than the tabulated value, then reject the null hypothesis.

(xi) If the computed value is less than the tabulated value, do not reject the null hypothesis.

The Chow test helps us in testing for parameter stability. Note that there are three limitations of the Chow test.

(i) We assume that the error variance σ^2 is constant throughout the sample. There is no difference in the error variance between the sub-samples.

(ii) The point of structural break is not known to us. We assume that point of structural change.

(iii) We cannot apply Chow test if there are more than one structural break.

8.8 PREDICTION

In Unit 5 we explained how prediction is made on the basis of simple regression model. We extend the same procedure to multiple regression models. As in the case of simple regression models, there are two types of prediction in multiple regression models.

If we predict an individual value of the dependent variable corresponding to particular values of the explanatory variables, we obtain the 'individual prediction'. When we predict the expected value of *Y* corresponding to particular values of the explanatory variables, it is called 'mean prediction'. The expected of *Y* in both the cases (individual prediction and mean prediction) is the same. The difference between mean and individual predictions lies in their variances.

8.8.1 Mean Prediction

Let

$$\mathrm{X}_0 = \begin{bmatrix} 1 \\ X_{02} \\ X_{03} \\ \vdots \\ X_{0k} \end{bmatrix} \quad (8.32)$$

be the vector of values of the X variables for which we wish to predict $\hat{Y}_0$.

The estimated multiple regression equation, in scalar form, is

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots \hat{\beta}_k X_{ki} + u_i \quad \text{... (8.33)}$$

which in matrix notation can be written compactly as

$$\hat{Y}_i = X_i' \hat{\beta} \quad \text{... (8.34)}$$

where

$$X_i' = [1 \;\; X_{2i} \; X_{3i} \ldots X_{ki}] \quad \text{... (8.35)}$$

and

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \text{... (8.36)}$$

Equation (8.34) is the mean predication of Y_i corresponding to given X_i'.

If X_i' is as given in (8.35), then (8.34) becomes

$$\left(\hat{Y}_i | X_0'\right) = X_0' \hat{\beta} \quad \text{... (8.37)}$$

where the values of x_0 are fixed. You should note that (8.36) gives an unbiased prediction of $E\left(\hat{Y}_i | X_0'\right)$, since $E\left(X_0' \hat{\beta}\right) = X_0' \hat{\beta}$.

Variance of Mean Prediction

The formula to estimate the variance of $\left(\hat{Y}_0 | X_0'\right)$ is as follows:

$$\text{var}\left(\hat{Y}_0 | X_0'\right) = \sigma^2 X_0' (X'X)^{-1} X_0 \quad \text{... (8.38)}$$

where σ^2 is the variance of u_i

X_0' are the X variables for which we wish to predict, and

since we do not know the error variance (σ^2), we replace it by its unbiased estimator $\hat{\sigma}^2$.

8.8.2 Individual Prediction

As mentioned earlier, expected value of individual prediction is the same as that of individual prediction, i.e., $\hat{Y}_i$. The variance of the individual prediction is

$$\text{var}(Y_0 | X_0) = \sigma^2 [1 + X'_0 (X'X)^{-1} X_0] \quad \text{... (8.39)}$$

where $\text{var}(Y_0 | X_0)$ stands for $E\left[Y_0 - \hat{Y}_0 \,|\, X\right]^2$. In practice we replace σ^2 by its unbiased estimator $\hat{\sigma}^2$.

Check Your Progress 4

1) Consider a Cobb-Douglas production. Write down the steps of testing the hypothesis that it exhibits constant returns to scale.

.....

.....

.....

.....

2) Write down the steps of carrying out Chow test.

.....

.....

.....

.....

3) Point out why individual prediction has higher variance than mean prediction.

.....

.....

.....

.....

8.9 LET US SUM UP

This unit described the assumptions of classical multiple regression that fortifies normality of error term also tested by Jarque-Bera Test (J-Test for Normality). The testing of hypothesis about individual coefficients is distinguished from the overall significance test in the unit. The unit also describes the testing of equality of two regression coefficients. Later the structural stability is tested using Chow test. The multiple regression is also used for prediction of dependent variables for given values of independent variables. Both individual and joint hypothesis testing is described in the unit. Various tests such as likelihood ratio (LR), Wald (W) and Lagrange Multiplier Test (LM) are explained in the unit

8.10 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) Refer to Sub-Section 8.2.1 and answer.

2) The Jarque-Bera test statistic is given at equation (8.2). Describe how thetest is carried out.

Check Your Progress 2

1) Refer to Sub-Section 8.3.1 and answer. Decide on the null and alternative hypotheses. Describe the staps you would follow.

2) Refer to Sub-Section 8.3.2 and answer.

Check Your Progress 3

1) It can be tested by F-test. See Section 8.4 for details.

2) Refer to Sub-Section 8.5 and answer.

Check Your Progress 4

1) We have explained in Sub-Section 8.6.1. Refer to it.

2) Refer to Sub-Section 8.7 and answer.

3) Refer to Sub-Section 8.8 and answer. It has the same logic as in the case of two variable models discussed in Section 5.7 of Unit 5.

UNIT 9 EXTENSION OF REGRESSION MODELS: DUMMY VARIABLE CASES*

Structure

9.0 OBJECTIVES

After reading this unit, you will be able to:

- define a qualitative or dummy variable;
- discuss the ANOVA model with a single dummy as exogenous variable;
- specify an ANCOVA model with one quantitative and one dummy variable;
- interpret the results of dummy variable regression models;
- differentiate between 'differential intercept coefficient' and 'differential slope coefficient;
- describe the concepts of 'concurrent, dissimilar and parallel' regression models that you encounter while considering 'differential slope dummies'; and
- explain how more than two dummies and interactive dummies can be formulated into a regression model.

9.1 INTRODUCTION

In real life situations, some variables are qualitative. Examples are gender, choices, nationality, etc. Such variables may be dichotomous or binary, i.e., with responses limited to two such as in 'yes' or 'no' situations. Or they may have more than two categorical responses. We need methods to include such variables in the regression model. In this unit, we consider some such cases. We limit this unit to consider regressions in which the dependent variable is quantified. You may note in passing that when the dependent variable itself is a dummy variable, we have to deal with them by models such as Probit or Logit. In such models, the

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi and Prof. B S Prakash, Indira Gandhi National Open University, New Delhi

OLS method of estimation does not apply. In this unit, we will not consider such cases. You will study about them in the course 'BECE 142: Applied Econometrics'.

In this unit, we consider only such cases in which the independent variable is a dummy variable. Qualitative variables are not straightaway quantified. By treating them as dummy variables we can make them quantified (or categorical). For instance, consider variables such as male or female, employed or unemployed, etc. These are quantifiable in the sense that by treating them as 1 if 'female, and 0 if 'male'. Similar examples could be 1 if yes and 0 if no; 1 if employed and 0 if unemployed, etc. In the above, we have converted a qualitative response into quantitative form. Thus, the qualitative variable is now quantified. Such regressions could be a simple regression, i.e., there is only one independent variable which is qualitative and treated as dummy variable. Or there could be two independent variables, one of which can be treated as dummy and the other is its covariant, i.e., there is a close relationship with the variable treated as dummy. For instance, pre-tax income of persons can be classified above a threshold level and treated as dummy variable, i.e., above or below the threshold level income with response taken as 1 or 0. Now, the post-tax income, which is a co-variant of pre-tax income, can be considered by its actual quantified value. There could be similar extension of situations where you have to consider multiple dummies and cases where you have to consider interactive dummies. The nature of such regressions, particularly for their inference or interpretational interest, is what we consider in the present unit.

9.2 THE CASE OF SINGLE DUMMY: ANOVA MODEL

We first consider a simple regression model with only one independent variable. Further, this independent variable is a dummy variable such as:

$$Y_i = \beta_1 + \beta_2 D_i + u_i \quad \ldots (9.1)$$

Here, we take Y as the annual expenditure on food and D_i as gender taking the values 0 if the person is male and 1 if female. The D_i's are thus fixed and hence non-stochastic. Now, if we assume that $u_i \sim N(0, \sigma^2)$, the OLS method can be applied to estimate the parameters in (9.1). If we do this, the mean food expenditure for males and females are respectively given by:

$$E(Yi \mid Di = 0) = \beta_1 + \beta_2(0) = \beta_1 \quad \ldots (9.2)$$

$$E(Y_i \mid D_i = 1) = \beta_1 + \beta_2 \quad \ldots (9.3)$$

Here, β_1 gives the average or mean food expenditure of males. It is the category for which the dummy variable is given the value 0. The slope coefficient β_2 tells us by how much the mean food expenditure of females differ from that of the mean food expenditure of males. Hence, $\beta_1 + \beta_2$ gives the mean food expenditure for females. In view of this, it is not correct to call β_2 as the slope coefficient since there is no continuous regression line here. Hence, β_2 is the 'differential

intercept coefficient'. It tells us by how much the value of intercept term differs between the two categories. A question that arises now is, what would have happened if we had interchanged the assignment of '0' between the two categories of males and females (i.e., if we had assigned the value '0' to females). You may note that, so long as we have only two categories as in the present instance, i.e., it is a case of simple regression with only one independent variable taken as a dummy variable D_i with the category of responses dichotomous or binary, it basically does not matter which category gets the value of 1 and which gets the value 0. However, some minor difference would be there. Let us see what this is.

The category to which we assign the value 0 is called as the base category. It is also called by alternative names such as reference or benchmark or the comparison category. In such an assignment, the intercept value represents the mean value of the category that gets the value 0 (which is males in our case above). What equation (9.3) tells us is, depending on such an assignment, the mean value of expenditure on food for females is to be obtained by adding the 'slope coefficient to the intercept value'. If the assignment of dummy is made the other way, i.e., females 0 and males 1, we see a change in the numerical value of the intercept term and its t value. Barring this, the R^2 value, the absolute value of the estimated dummy variable coefficient and its standard error, will remain the same. Let us see this with the help of an example for better understanding.

Consider the data on 'expenditure on food' and income for males and females as in Table 9.1. The data are averages based on the actual number of people (who are in thousands) in different age groups. We first construct Table 9.2 from the data in Table 9.1 as below.

Table 9.1: Data on Income and Food Expenditure by Gender

(Figures in $)

Age	Food Expenditure (female)	Income (female)	Food Expenditure (male)	Income (male)
< 25	1983	11557	2230	11589
25-34	2987	29387	3757	33328
35-44	2993	31463	3821	36151
45-54	3156	29554	3291	35448
55-64	2706	25137	3429	32988
> 65	2217	14952	2533	20437

Source: Table 6-1, Chapter 6, Gujarati.

Table 9.2: Food Expenditure in Relation to Income and Gender

Observation	Food Expenditure ($)	Income ($)	Gender
1	1983	11557	1
2	2987	29387	1
3	2993	31463	1
4	3156	29554	1
5	2706	25137	1
6	2217	14952	1
7	2230	11589	0
8	3757	33328	0
9	3821	36151	0
10	3291	35448	0
11	3429	32988	0
12	2533	20437	0

Source: Table 6-2, Chapter 6, Gujarati.

Results of food expenditure regressed on the gender dummy variable (without taking into account the income variable at this stage) presents the following results.

$$\widehat{Y}_i = 3176.833 - 503.1667\ D_i$$

$$\text{se} = (233.0446) \quad (329.5749)$$

$$t = (13.6318) \quad (-1.5267) \quad R^2 = 0.1890$$

The results show that the mean expenditure of males is 3177 $ and that of females is (3177 – 503 = 2674 $). The estimated D_i is not statistically significant (since its t value is only –1.53). This means that the difference in the food expenditure between gender is not statistically significant. Recall that we have assigned the value '0' to males. Hence, the intercept value represents the mean value for males. In this assignment, to get the mean value of food expenditure of females, we add the value of the coefficient of the dummy variable to the intercept value. Now, let us re-assign the value '0' to females and '1' to males. The regression results that we get are the following:

$$\widehat{Y}_i = 2673.667 + 503.1667\ D_i$$
$$\text{se} = (233.0446) \quad (329.5749)$$
$$t = (11.4227) \quad (-1.5267) \quad R^2 = 0.1890$$

Thus, we notice that the mean food consumption expenditures of the two genders have remained the same. The R^2 value is also the same. The absolute value of the dummy variable coefficient and their standard errors are also the same. The only change is in the numerical value of the intercept term and its t value.

Another question that we may get is: since we have two categories, male and female, can we assign two dummies to them? This means we consider the model as:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_i + u_i \quad \ldots (9.4)$$

where Y is expenditure on food, $D_2 = 1$ for female and 0 for male and $D_3 = 1$ for male and 0 for female. Essentially, we are trying to see whether we can assign two dummies for male and female separately? The answer is 'no'. To know the reason for this, consider the data for a sample of two females and three males, for which the data matrix is as in Table 9.3. We see that $D_2 = 1 - D_3$ or $D_3 = 1 - D_2$. This is a situation of perfect collinearity. Hence, we must always use only one dummy variable if a qualitative variable has two categories, such as the gender here.

Table 9.3: Data Matrix for the Equation

Gender	Intercept	D_2	D_3
Male Y_1	1	0	1
Male Y_2	1	0	1
Female Y_3	1	1	0
Male Y_4	1	0	1
Female Y_5	1	1	0

A more general rule is: if a model has the common intercept β_1, and the qualitative variable has m categories, then we must introduce only $(m - 1)$ dummy variables. If we do not do this, we get into a problem of estimation called as the 'dummy variable trap'. Finally, note that when we have a simple regression model with only one dummy variable as considered here, the model considered is also called as the ANOVA model. This is because there is no second variable from which we are seeking to know the impact or variability on the dependent variable. When we have this, we get what we call as an ANCOVA model. We take up such a case in the next section.

9.3 ANALYSIS OF COVARIANCE (ANCOVA) MODEL

In economic analysis, it is common to have among explanatory variables some of which are qualitative and some others quantitative. Such models are called as Analysis-of-Covariance (ANCOVA) models. Here, we shall consider a model that has both a quantitative and a dummy variable among the regressors. In general, regression models containing a combination of quantitative and qualitative variables are called ANCOVA models. Here, the quantitative variables are called covariates or control variables. ANCOVA models are an extension of the ANOVA models. They provide a method of statistically controlling the effects of covariates (i.e., a quantitative explanatory variable) in a model that includes both the type of variables with the qualitative variable treated as a dummy variable. The quantitative variable considered is usually a covariate in the sense that it bears close association with the main variable. Because of this, exclusion of covariates from a model results in model specification error. In the example considered above, we regressed 'food expenditure' on only gender dummy [$Y_i = \beta_1 + \beta_2 D_i + u_i$]. Now, let us consider another variable, 'income after taxes', i.e., disposable income (a covariate of food expenditure) as an explanatory variable (X_i). The model now is

$$Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i \quad \ldots (9.5)$$

where Y = expenditure on food ($), X = after tax income ($), D = 1 for female and = 0 for male. Let us now consider, for better appreciation, the result for the regression in equation (9.5) obtained from the data in Table 9.2 as follows:

$$\hat{Y}_c = 1506.244 - 228.9868 D_i + 0.0589 X_i$$
t = (8.0115) (–2.1388) (9.6417)
R^2 = 0.9284

The dummy variable coefficient is statistically significant. Therefore, we reject the null hypothesis that there is no difference in the average value of expenditure on food for male and female. In other words, we conclude that gender has a significant impact on consumption or food expenditure. Note that this difference in consumption expenditure is inferred holding the effect of after-tax income constant. Likewise, holding the gender differences constant, the after tax income coefficient is significant. The slope coefficient for 'after tax income' indicates that the mean food expenditure [i.e., the marginal propensity to consume (MPC)] increases by 6 cents for every additional dollar of increase in the disposable income. Note that since we have taken '0' for males, the intercept term relates to the MPC for males. For female MPC, we have to add the intercept value to the coefficient of gender dummy (i.e., 1506.2 – 228.9 = 1277.3). Thus, the equations for the MPC of females and males can be respectively written as:

Mean food expenditure for females: $\hat{Y}_i = 1277.2574 + 0.0589 X_i$

Mean food expenditure for males: $\hat{Y}_i = 1506.2440 + 0.0589 X_i$

Since the MPC or the slope is same for both the gender, the two regressions are parallel as in Fig. 9.1 below.

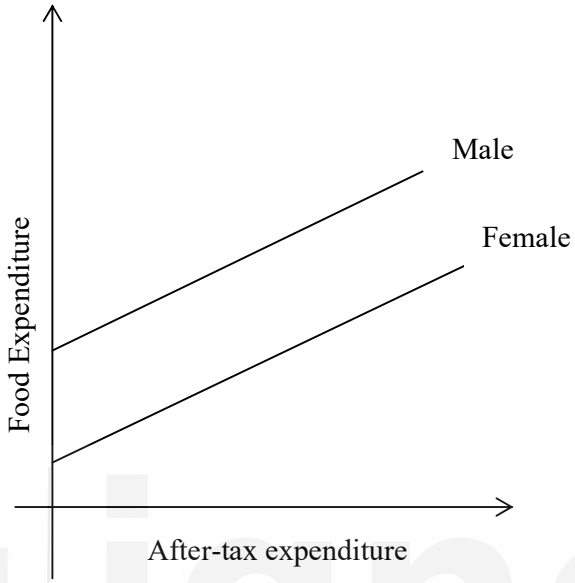


Fig. 9.1 Mean Food Expenditure for Male and Female

The model signifies the role and the impact of both the type of variables (quantitative and qualitative) in explaining a dependent variable. Specifically, in the example considered, the after tax expenditure is seen to affect the food expenditure of both males and females.

Check Your Progress 1 [answer questions in about 50-100 words]

1) Define a qualitative variable.

.....

.....

.....

.....

.....

2) Specify a regression model with a single dummy variable. Mention its features from the point of view of interpretation of estimated coefficients.

.....

.....

.....

.....

.....

3) What happens if the base value is reassigned for the dummy variable, say gender, in a simple regression model as in equation (9.1)?

.....

.....

.....

.....

.....

4) What is meant by 'dummy variable trap'? How do we avoid it?

.....

.....

.....

.....

.....

5) Distinguish between an ANOVA model and an ANCOVA.

.....

.....

.....

.....

.....

6) What is an advantage of ANCOVA model? What is a consequence of omitting the inclusion of a covariant in an ANOVA model?

.....

.....

.....

.....

.....

7) Specify the general form of an ANCOVA model with one qualitative and one quantitative variable. What does the slope oefficient for the quantitative variable considerd indicate in general?

.....

.....

.....

.....

.....

9.4 COMPARISON BETWEEN TWO REGRESSION MODELS

In the example considered above, i.e., for both the ANOVA and the ANCOVA models, we saw that the slope coefficients were same but the intercepts were different. This raises the question on whether the slopes too could be different? How do we formulate the model if our interest is to test for the difference in the slope coefficients too? In order to capture this, we introduce a 'slope drifter'. For the example of consumption expenditure for male or female considered above, let us now proceed to compare the difference in the consumption expenditure by gender by specifying the model with dummies as follows:

$$Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i \quad \ldots (9.6)$$

Note that the additional variable added is $D_i X_i$ which is in multiplicative or interactive form. In (9.6), we have taken $D_i = 0$ for males and $D_i = 1$ for females. Now, the 'mean food expenditure' for males is given by:

$$E(Y_i \mid D_i = 0, X_i) = \beta_1 + \beta_3 X_i \quad \ldots (9.7)$$

{since $D_i = 0$}

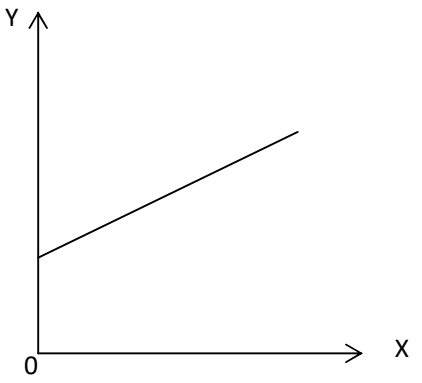
The 'mean food expenditure' for females is given by:

$$E(Y_i \mid D_i = 1, X_i) = \beta_1 + \beta_2 D_i + (\beta_3 + \beta_4 D_i) X_i$$

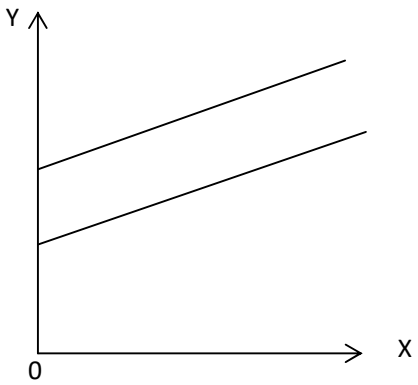
$$= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) X_i \quad \ldots (9.8)$$

{since $D_i = 1$}

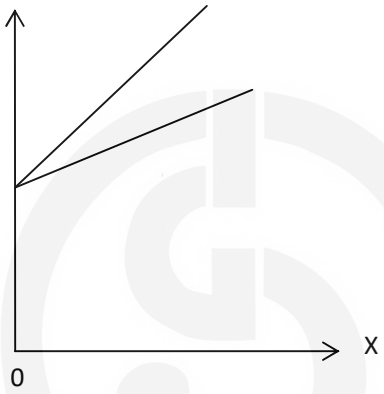
In equation (9.8), $(\beta_1 + \beta_2)$ gives the mean value of Y for the category that receives the dummy value of 1 when X is zero. And, $(\beta_3 + \beta_4)$ gives the slope co-efficient of the income variable for the category that receives the dummy value of 1. Note that the introduction of the dummy variable in the 'additive form' enables us to distinguish between the intercept terms of the two groups. Likewise, the introduction of the dummy variable in the interactive (or multiplicative) form (i.e., $D_i X_i$) enables us to differentiate between the slope coefficients (or terms) of the two groups. Depending on the statistical significance of the differential intercept coefficient, β_2, and the differential slope coefficient, β_4, we can infer whether the female and male food expenditure functions differ in their intercept values, or their slope values, or both. There can be four possibilities as shown in Fig. 9.2. Fig. 9.2 (a) shows that there is no difference in intercept or the slope coefficient of the two food expenditure regressions. Such regression equations are called 'Coincident Regressions'.



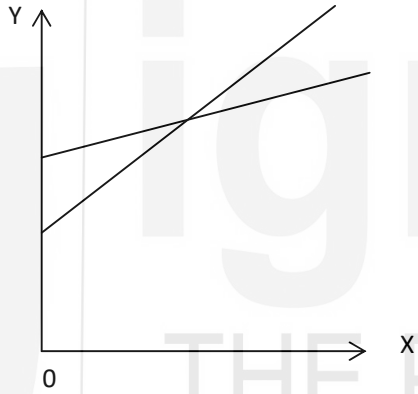
(a) Coincident Regressions



(b) Parallel Regressions



(c) Concurrent Regressions



(d) Dissimilar Regressions

Fig 9.2 Comparison of Regression Equations

Fig. 9.2 (b) shows that the two slope coefficients are the same but intercepts are different. Such regressions are referred to as 'Parallel Regressions'. Fig. 9.2 (c) shows that the two regressions have the same intercepts but

different slopes. Such regressions are referred as 'Concurrent Regressions'. Fig. 9.2 (d) shows that the two intercepts and the two slope coefficients are both different. Such regressions are called 'Dissimilar Regressions'.

9.5 MULTIPLE DUMMIES AND INTERRACTIVE DUMMIES

We often might require to consider more than one dummy variables. Besides, there could be cases where we might be interested in seeing for the impact of dummy variable interactions. Let us consider a case as given below.

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i \quad \ldots (9.7)$$

where Y is income, X is education measured in number of years of schooling, D_2 is gender (0 if male, 1 if female), D_3 is if in reserved segment or group (e.g. SC/ST/OBC) taking the value 0 if 'not in reserved segment', i.e., in general segment and 1 if 'in reserved segment'. Here, gender (D_2) and reservation (D_3) are qualitative variables and X is quantitative variable. In this formulation (for example, equation 9.7) we have made an implicit assumption that the differential effect of gender is constant across the two segments of reservation. We have likewise assumed that the differential effect of reservation is constant across the two genders. This means if the average income is higher for males than for females, it is so whether the person is in the general segment or in the reservation segment. Likewise, it is assumed here that if the average income is different between the two reservation segments, it is so irrespective of gender. However, in many cases, such assumptions may not be tenable. This means, there could be interaction between gender and reservation dummies. In other words, their effect on average income may not be simply additive as in (9.7) but could be multiplicative. If we wish to consider for this interactive effect, we must specify the model as follows:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i}\, D_{3i}) + \beta_5 X_i + u_i \quad \ldots (9.8)$$

In equation (9.8), the dummy variable $D_{2i}D_{3i}$ is called as 'interactive or interaction dummy'. It represents the joint or simultaneous effect of two qualitative variables. Taking expectation on both sides of equation (9.8), i.e., by considering the average effect on income across gender and reservation, we get:

$$E(Yi \mid D_{2i}=1, D_{3i}=1, X_i) = \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 X_i \quad \ldots (9.9)$$

Equation (9.9) is the average income function for female reserved category workers where β_2 is the differential effect of being female, β_3 is the differential effect of being in the reserved segment and β_4 is the interactive effect of being both a female and in reserved segment. Depending on the statistical significance of various dummies, we need to make relevant inferences. The specification can easily be generalized for more than one quantitative variable and more than two qualitative variables.

Check Your Progress 2 [answer questions within the given space in about 90-100 words]

1) What is meant by a 'slope drifter'? When is it introduced and for what use? Specify a general model with such a 'slope drifter' and comment on the additional variable introduced.

.....

.....

.....

.....

.....

2) Differentiate between the four type of regressions that we might get when considering a model of the type in equation (9.6) with two slope drifters β_2 and β_4 as therein.

.....

.....

.....

.....

.....

3) List the four types of regression models, with dummy variables to accommodate different cases or situations, as we have considered in this unit. Specify their difference by their name and features.

.....

.....

.....

.....

.....

9.6 LET US SUM UP

This unit makes a distinction between qualitative and quantitative variables. It has considered three types of models in which the focus is kept on inclusion of qualitative variables in the regression models. The first of such models is considered is a simple regression model. In this, we have considered only one dummy variable, as an independent variable, on the RHS of the regression equation. This equation is of the form: $Y_i = \beta_1 + \beta_2 D_i + u_i$. Analysis in this form is called as ANOVA. Quite often, we would be committing a specification bias if we consider the regression model in this form. This happens because the variable Y_i will be clearly related to a variable X_i which is a quantitative variable. To accommodate this, we considered the second type of model in which we included a co-variant (X_i) into the regression equation: $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$. Analysis in this form is called as ANCOVA. In both these type of models, our focus was only on observing the significance of difference in the intercepts. But in practice, we do encounter a number of situations in which not only the intercept, but the slope too could vary between categories. To allow for this kind of situation, we considered a third type of model in which we accommodated for the interactive effect of the 'dummy variable with the quantitative variable', i.e., D_iX_i. The regression model considered for this kind of an analysis is of the form: $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$. In this situation, we noted that we could come across four possibilities viz. coincidental, parallel, concurrent and dissimilar regressions. We have finally considered the case where a regression model may have to be formulated to accommodate more than one qualitative

variable and a case where we might be interested in examining for the interactive effect of the two qualitative variables. For this, we considered models such as $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i} D_{3i}) + \beta_5 X_i + u_i$.

9.7 ANSWERS/ HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) A qualitative variable is one which has a categorical response such as yes/no or employed/unemployed or male/female. If the response is limited to two, as in these cases, it is called as a dichotomous variable. The responses can be more than two. But they may be classified as 1, 2, 3, …….. Such responses are unambiguous or categorical. Hence, a qualitative variable is also called as dummy variable or categorical variable.

2) The model in this case can be $Y_i = \beta_1 + \beta_2 D_i + u_i$. We are considering the dependent variable Y_i as quantitative variable. The D_i's are thus fixed and hence non-stochastic. D_i is taken a dichotomous, i.e., it takes the values 0 and 1. In such cases, the factor or entity which is assigned the value 0, is called as the base category. The estimated value of the mean of Y_i, given $D_i = 0$, is given by β_1. Here, β_2 is not strictly the slope coefficient but is the 'differential intercept coefficient'. The estimated value of the mean of Y_i, given $D_i = 1$, is given by $\beta_1 + \beta_2$.

3) The mean value of Y_i for the two gender classes, the R^2 value, the absolute value of the estimated dummy variable coefficient and the standard errors will be the same. The numerical value of the intercept term and its t value will change.

4) The number of responses to the dummy variable is called as 'categories' of response. If the dummy variable refers to gender of the respondent, there are two categories of response viz. male and female. If we assign two separate dummies in such cases, we encounter a situation of perfect collinearity. Hence, we will not get unique estimates or one of the two parameters is not estimable. This situation is called as 'dummy variable trap'. To avoid this situation, the general rule is if we have m categories, we limit the number of dummies to '$m - 1$'. The models should also have a common intercept β_1.

5) If the regression model considered has only one independent variable in general, and that variable is a dummy variable as considered here in particular, then the variation or the sources of variability that is sought to be identified for the dependent variable is limited to that one variable. In such cases, the regression model considered is called as an ANOVA model. If the independent variables considered are two, with one considered as dummy variable, and the other variable considered is related to the dummy variable, then such models are called as ANCOVA model.

In other words, regression models in which some independent variables are qualitative and some others are quantitative, are called as ANCOVA models.

6) The advantage is that ANCOVA models provide a method of statistically controlling the effects of covariates. The consequence of excluding a covariant from being included in the model is that the model suffers from 'specification error'. The consequence of committing specification errors are that the ideal assumptions required for the OLS estimators to be efficient are violated. Consequently, they lose out on their efficiency properties.

7) The general form of the model is like: $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$. The slope coefficient indicates the rate of increase (or decrease) in the 'marginal propensity to consume (MPC)'. This is when the dependent variable *Y* relates to a consumption variable like expenditure on food and the quantitative independent variable is like disposable income as considered here.

Check Your Progress 2

1) In regression models with one intercept and one slope coefficients, our interest might be to test to know whether: (i) the intercept terms are statistically different and (ii) the slope coefficients are statistically different? For investigating the second question, we need to introduce what is called as a 'slope drifter'. The model specified with such a drifter would be like: $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$.The additional variable introduced here is *DiXi*. It is a multiplicative variable in the interactive form. Here β_2 and β_4 are the two slope drifters which helps us infer for the statistical difference in the intercept values and the slope values respectively.

2) We get a 'coincident regression' when there is no difference both in intercept as well as the slope. We get a 'parallel regression' when the two intercept terms are different but the two slope coefficients are the same. We get a 'concurrent regression' when the two regressions have the same intercept but different slopes. We get two 'dissimilar regressions' when both the intercept terms and the slope coefficients are different.

3) (i) $Y_i = \beta_1 + \beta_2 D_i + u_i$. (ii) $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$. (iii) $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$. (iv) $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i}\, D_{3i}) + \beta_5 X_i + u_i$. The first is the ANOVA model in which we have considered only one single dummy variable as the independent variable. The second is the ANCOVA model in which we have considered one qualitative dummy variable and another quantitative exogenous variable related to the dummy variable, the omission of which would lead to a 'specification bias'. The third involves an interactive variable $(D_i X_i)$ in which we try to see whether both the slopes and the intercept coefficients differ. In this, there is a possibility of getting four different type of regressions viz. coincident, parallel, concurrent and dissimilar regressions. The fourth situation considered involves a interactive dummy variable like: $Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (D_{2i}\, D_{3i}) + \beta_5 X_i + u_i$.

UNIT 10 MULTICOLLINEARITY*

Structure

10.0 OBJECTIVES

After going through this unit, you should be able to

- explain the concept of multicollinearity in a regression model;
- comprehend the difference between the near and perfect multicollinearity;
- describe the consequences of multicollinearity;
- [1]explain how multicollinearity can be detected; and
- describe the remedial measures of multicollinearity; and
- explain the concept of ridge regression.

10.1 INTRODUCTION

The classical linear regression model assumes that there is no perfect multicollinearity. Multicollinearity means the presence of high correlation between two or more explanatory variables in a multiple regression model.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

Absence of multicollinearity implies that there is no exact linear relationship among the explanatory variables. The assumption of no perfect multicollinearity is very crucial to a regression model since the presence of perfect multicollinearity has serious consequences on the regression model. We will discuss about the consequences, detection methods, and remedial measures for multicollinearity in this Unit.

10.2 TYPES OF MULTICOLLINEARITY

Multicollinearity could be of two types: (i) perfect multicollinearity, and (ii) imperfect multicollinearity. Remember that the division is according to the degree or extent of relationship between the explanatory variables. The distinction is made because of the nature of the problem they pose. We describe both types of multicollinearity below.

10.2.1 Perfect Multicollinearity

In the case of perfect multicollinearity, the explanatory variables are perfectly correlated with each other. It implies the coefficient of correlation between the explanatory variables is 1. For instance, suppose want to derive the demand curve for a good Y. We assume that quantity demanded (Y) is a function of price (X_2) and income (X_3). In symbols,

$Y = f(X_2, X_3)$ where X_2 is price of good Y and X_3 is the weekly consumer income.

Let us consider the following regression model (population regression function):

$$Y_i = A_1 + A_2X_{2i} + A_3X_{3i} + u_i \quad \ldots (10.1)$$

In the above equation, suppose

A_2 is < 0. This implies that prices are inversely related do demand.

$A_3 > 0$. This indicates that as income increases, demand for the good increases.

Suppose there is a perfect relationship between X_2 and X_3 such that

$$X_{3i} = 300 - 2X_{2i} \quad \ldots (10.2)$$

In the above case, if we regress X_3 on X_2 we obtain the coefficient of determination $R^2 = 1$.

If we substitute the value of X_3 from equation (10.2), we obtain

$$\begin{aligned} Y_i &= A_1 + A_2X_{2i} + A_3(300 - 2X_{2i}) + u_i \\ &= A_1 + A_2X_{2i} + 300A_3 - 2A_3X_{2i} + u_i \\ &= (A_1 + 300A_3) + (A_2 - 2A_3)X_{2i} + u_i \quad \ldots (10.3) \end{aligned}$$

Let $C_1 = (A_1 + 300A_3)$ and $C_2 = (A_2 - 2A_3)$. Then equation (10.3) can be written as:

$$Y_i = C_1 + C_2X_{2i} + u_i \quad \text{....(10.4)}$$

Thus if we estimate the regression model given at (10.4), we obtain estimators for C_1 and C_2. We do not obtain unique estimators for A_1, A_2 and A_3.

As a result, in the case of perfect linear relationship or perfect multicollinearity among explanatory variables, we cannot obtain unique estimators of all the parameters. Since we cannot obtain their unique estimates, we cannot draw any statistical inferences (hypothesis testing) about them. Thus, in case of perfect multicollinearity, estimation and hypothesis testing of individual regression coefficients in a multiple regression are not possible.

10.2.2 Near or Imperfect Multicollinearity

In the previous section, the presence of perfect multicollinearity indicated that we do not get unique estimators for all the parameters in the model. In practice, we do not encounter perfect multicollinearity. We usually encounter near or very high multicollinearity. In this case the explanatory variables are approximately linearity related.

High collinearity refers to the case of "near" or "imperfect" multicollinearity. Thus, when we refer to the problem of multicollinearity we usually mean "imperfect multicollinearity''

Let us consider the same demand function of good Y. In this case we however assume that there is imperfect multicollinearity between the explanatory variables (in order to distinguish it from the earlier case, we have changed the parameter notations). The following is the population regression function:

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + u_i \quad \text{....(10.5)}$$

Equation (10.5) refers to the case when two or more explanatory variables are not exactly linear. For the above regression model, we may obtain an estimated regression equation as follows:

Equation (10.5):	$\hat{Y}_i = 145.37$	$-$	$2.7975X_{2i}$	$-$ $0.3191X_{3i}$
Standard Error:	(120.06)		(0.8122)	(0.4003)
t-ratio:	(1.2107)		(–3.4444)	(–0.7971)

$R^2 = 0.97778$ ... (10.6)

Since the explanatory variables are not exactly related, we can find estimates for the parameters. In this case, regression can be estimated unlike the first case of prefect multicollinearity. It does not mean that there is no problem with our estimators if there is imperfect multicollinearity. We discuss the consequences of multicollinearity in the next section.

Check Your Progress 1

1) What is meant by perfect multicollinearity?

.....

.....

.....

.....

.....

2) What do you understand by imperfect multicollinearity?

.....

.....

.....

.....

.....

3) Explain why it is not possible to estimate a multiple regression model in the presence of perfect multicollinearity.

.....

.....

.....

.....

.....

10.3 CONSEQUENCES OF MULTICOLLINEARITY

We know from Unit 4 that the ordinary least squares (OLS) estimators are the Best Linear Unbiased Estimators (BLUE). It implies they have the minimum variance in the class of all linear unbiased estimators. In the case of imperfect multicollinearity, the OLS estimators still remain BLUE. Then what is the problem? In the presence of multicollinearity, there is an increase in the variance and standard error of the coefficients. As a result, very few estimators are statistically significant.

Some more consequences of multicollinearity are given below.

(a) The explanatory variables may not be linearly related in the population (i.e., in the population regression function), but they could be related in a particular sample. Thus multicollinearity is a sample problem.

(b) Near or high multicollinearity results in large variances and standard errors of OLS estimators. As a result, it becomes difficult to estimate true value of the estimator.

(c) Multicollinearity results in wider confidence intervals. The standard errors associated with the partial slope coefficients are higher. Therefore, it results in wider confidence intervals.

$$P_r\left[b_2 - t_{\alpha/2}SE(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2}SE(b_2)\right] = 1 - \alpha \quad \ldots.(10.7)$$

Since the values of standard errors have increased the interval reflected in expression in (10.7) has widened.

(d) Insignificant t ratios: As pointed out above, standard errors of the estimators increase due to multicollinearity. The t-ratio is given as $= \frac{b_2}{SE(b_2)}$. Therefore, the t-ratio is very small. Thus we tend to accept (or do not reject) the null hypothesis and tend to conclude that the variable has no effect on the dependent variable.

(e) A high R^2 and few significant t-ratios: In equation (10.6) we notice that the R^2 is very high, about 98% or 0.98. The t-ratios of both the explanatory variables are not statistically significant. Only the price variable slope coefficient has significant t-value. However, using F-test while testing overall significance $H_0: R^2 = 0$, we reject the null hypotheses. Thus there is some discrepancy between the results of the F-test and the t-test.

(f) The OLS estimators are mainly partial slope coefficients and their standard errors become very sensitive to small changes in the data. If there is a small change in data, the regression results change substantially.

(g) Wrong signs of regression coefficients: It is a very prominent impact of the presence of multicollinearity. In the case of the example given at equation (10.6) we find that the coefficient of the variable income is negative. The income variable has a 'wrong' sign as economic theory suggests that income effect is positive unless the commodity concerned is an inferior good.

10.4 DETECTION OF MULTICOLLINEARITY

In the previous section we pointed out the consequences of multicollinearity. Now let us discuss how multicollinearity can be detected.

(h) High R^2 and Few Significant t-ratios

This is the classic symptom of multicollinearity. If R^2 is high (greater than 0.8), the null hypothesis that the partial slope coefficients are jointly or simultaneously equal to zero $\left[H_0 : \beta_2 = \beta_3 = 0\right]$ is rejected in most cases (on the basis of F-test). But the individual t-tests will reflect that none or very few partial slope coefficients are statistically different from zero. This suggests very few slope coefficients are statistically significant.

(ii) High Pair-wise Correlations among Explanatory Variables

Due to high correlation among the independent variables, the estimated regression coefficients have high standard errors. But this is not necessarily true as demonstrated below. Even low correlation among the independent variables can lead to the problem of multicollinearity.

Let r_{23}, r_{24} and r_{34} represent the pair-wise correlation coefficients between X_2 and X_3 and X_4 respectively. Suppose $r_{23} = 0.90$, reflecting high collinearity between X_2 and X_3. Let us consider partial correlation coefficient $r_{23.4}$ that indicates correlation between X_2 and X_3 (while keeping the influence of X_4 constant). Suppose we find that $r_{23.4} = 0.43$. It indicates that partial correlation between X_2 and X_3 is low reflecting the absence of high collinearity. Therefore, pair-wise correlation coefficient when replaced by partial correlation coefficients does not indicate the presence of multicollinearity. Suppose the true population regression is given by equation (10.8)

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_2 X_{3i} + \beta_4 X_{4i} + u_i \quad \ldots (10.8)$$

Suppose the explanatory variables are perfectly correlated with each other as shown in equation (10.9) below

$$X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i} \quad \ldots (10.9)$$

X_4 is an exact linear combination of X_2 and X_3

If we estimate the coefficient of determination by regressing X_4 on X_2 and X_3, we find that

$$R^2_{4.23} = \frac{r^2_{42} + r^2_{43} - 2r_{42}\, r_{43}\, r_{23}}{1 - r^2_{23}} \quad \ldots (10.10)$$

Suppose, $r_{42} = 0.5, r_{43} = 0.5, r_{23} = -0.5$. If we substitute these values in equation (10.10), we find that $R^2_{4.23} = 1$. An implication of the above is that all the correlation coefficients (among explanatory variables) are not very high but still there is perfect multicollinearity.

(iii) Subsidiary or Auxiliary Regressions

Suppose one explanatory variable is regressed on each of the remaining variables and the corresponding R^2 is computed. Each of these regressions is referred to as subsidiary or auxiliary regression. For example, in a regression model with seven explanatory variables, we regress X_1 on X_2, X_3, X_4, X_5, X_6 and X_7 and find out the R^2_1. Similarly, we can regress X_2 on X_1, X_3, X_4, X_5, X_6 and X_7 and find out the R^2_2. By examining the auxiliary regression models we can find out the possibility

of multicollinearity. We take the rule of thumb that multicollinearity may be troublesome if R_i^2 obtained from auxiliary regression is greater than overall R^2 of the regression model.

A limitation of this method is that we have to compute R_i^2 several times, which is cumbersome and time consuming.

(iv) Variance Inflation Factor (VIF)

Another indicator of multicollinearity is the variance inflation factor (VIF). The R_i^2 obtained from auxiliary regressions may not be a reliable indicator of collinearity. In VIF method we modify the formula of variance of the estimators as follows; (b_2) and (b_3)

$$\text{var}\,(b_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-R_2^2)} = \frac{\sigma^2}{\sum X_{2i}^2}.\left(\frac{1}{1-R_2^2}\right) \quad \text{... (10.11)}$$

In equation (10.11), you should note that R_2^2 is the auxiliary regression discussed earlier.

Compare the variance of b_2 given in equation (10.11) with the usual formula for variance of an estimator given in Unit 4. We find that

$$var(b_2) = \frac{\sigma^2}{\sum x_{2i}^2}\ VIF \quad \text{... (10.12)}$$

where VIF $= \left(\frac{1}{1-R_2^2}\right)$

Similarly, $var(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}(VIF)$

Note that as R_i^2 increases the VIF also increases. This inflates the variance and hence standard errors of b_2 and b_3

If $R_i^2 = 0$, $VIF = 1 \Rightarrow V(b_2) = \frac{\sigma^2}{\sum x_{2i}^2}$ and $V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}$

Therefore, there is no collinearity.

On the other hand,

if $R_i^2 = 1$, $VIF = \infty \Rightarrow V(b_2) \to \infty, V(b_3) \to \infty$

If R_i^2 is high, however $V(b_2)$ tends to ∞.

Note that $var(b_2)$ depends not only on R_i^2, but also on σ^2 and $\sum x_{2i}^2$. It is possible that R_i^2 is high (say, 0.91) but $var(b_2)$ could be lower due to low σ^2 or high $\sum x_{2i}^2$. Thus $V(b_2)$ is still lower resulting in high *t* value. Thus R_i^2 obtained from auxiliary regression is only a superficial indicator of multicollinearity.

Check Your Progress 2

1) Bring out four important consequences of multicollinearity.

.....

.....

.....

.....

.....

2) Explain how multicollinearity can be detected using partial correlations.

.....

.....

.....

.....

.....

3) Describe the method of detection of multicollinearity using the variance inflation factor (VIF).

.....

.....

.....

.....

.....

10.5 REMEDIAL MEASURES OF MULTICOLLINEARITY

Multicollinearity may not necessarily be an "evil'' if the goal of the study is to forecast the mean value of the dependent variable. If the collinearity between the explanatory variables is expected to continue in future, then the population regression function can be used to predict the relationship between the dependent variable Y and other collinear explanatory variables.

However, if in some other sample, the degree of collinearity between the two variables is not that strong the forecast based on the given Regression is of little use.

On the other hand, if the objective of the study is not only prediction but also reliable estimations of the individual parameters of the chosen model then serious collinearity may be bad, since multicollinearity results in large standard errors of estimators and therefore widens confidence interval. Thus, resulting in accepting null hypotheses in most cases. If the objective of the study is to estimate a group

of coefficients (i.e., sum or difference of two coefficients) then this is possible even in presence of multicollinearity. In such a case multicollinearity may not be a problem.

$$Y_i = C_1 + C_2 X_{2i} + u_i \qquad \ldots(10.13)$$

$$C_1 = A_1 + 300A_3, \qquad C_2 = A_2 - 2A_3$$

Running the above regression in equation (10.2), as presented in earlier section 10.2, one can easily estimate C_2 by using OLS method, although neither A_2 nor A_3 can be estimated individually. There can be situation when in spite of inflated S.E., the individual coefficients happened to be numerically significant since the true value itself is so large even or estimate on the downside still shows up a significant test.

Certain remedies prescribed for reducing the severity of collinearity problem which can be listed as OLS estimators can still retain BLUE property despite of near collinearity. Further, one or more regression coefficients can e individually statistically significant or some of them with wrong signs.

10.5.1 Dropping a Variable from the Model

The simplest solution may be to drop one or more of the collinear variables. However, dropping a variable from the model may lead to model specification error. In other words, when we estimate the model without the excluded variable, the estimated parameters of the reduced model may turn out to be biased. Therefore, the best practical advice is not to drop a variable from a model that is theoretically sound. A variable which has *t* value of its coefficient greater than 1, then than variable should not be dropped as it will result in a decrease in $\bar{R}^2$.

10.5.2 Acquiring Additional Data or New Sample

Acquiring additional data implies increasing the sample size. This is likely to reduce the severity of the multicollinearity problem. As we know from equation (10.11),

$$\text{var}\,(b_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - R_2^2)}$$

Given σ^2 and R_2^2, if the sample size of X_2 increases, there is an increase in $\sum x_{2i}^2$. It will lead a decrease in var (b_2) and its standard error.

10.5.3 Re-Specification of the Model

It is possible that some important variables are omitted from the model. The functional form of the model may also be incorrect. Therefore, there is a need of looking into the specification of the model. Many times, taking log form of a model leads to solving the problem of multicollinearity.

10.5.4 Prior Information about Certain Parameters

Estimated values of certain parameters are available in existing studies. These values can be used as prior information. These values give us some tentative idea on the plausible value of the parameters.

10.5.5 Transformation of Variables

Transformation of the variables would minimize the problem of collinearity.

10.5.6 Ridge Regression

The ridge regressions are another method of resolving the problem of multicollinearity. In the ridge regression, the first step is to standardize the variables both dependent and independent by subtracting the respective means and dividing by their standard deviations. This mainly implies that the main regression is run by transforming both dependent and explanatory variables into the standardized values.

It is observed that in the presence of multicollinearity, the value of variance inflation factor is substantially high. This is mainly due to a high value of coefficient of determination. The ridge regression is applied when the regression equations are in the form of matrix involving large number of explanatory variables.

The ridge regression proceeds by adding a small value, k, to the diagonal elements of the correlation matrix. The reason that the diagonal of ones in the correlation matrix could be considered as a ridge, this is the reason such regression is referred as ridge regression.

10.5.7 Other Remedial Measures

There are several other Remedies suggested such as combining time series and cross-sectional data, factor or principal component analysis and ridge regressions.

Polynomial Regression Models

Let us consider total cost of production (TC) as a function of output as well as marginal cost (MC) and Average Cost (AC)

$$Y_i = \beta_1 + \beta_2 X_1 + \beta_3 X_i^2 + \beta_4 X_i^3 \qquad \text{…….(10.12)}$$

The cost function is defined as Cubic function for cost as a third-degree polynomial of variable X. This model in equation (10.12) is linear in parameters β^s, therefore satisfy assumption of CLRM of linear Regression Model and can be estimated by usual OLS method. However, one needs to worry about problem of collinearity since it is not linear in variables and at the same time X^2 and X^3 are non-linear function of X and do not violate the assumptions of no perfect collinearity i.e., no perfect linear relationship between variables. The estimated results are presented in equation (10.13).

$$\hat{Y}_i = 141.7667 + 63.4776X_i - 12.9615X_i^2 + 0.9396\ X_i^3 \qquad \text{….(10.13)}$$

Se (6.3753) (4.7786) (0.9857) (0.0591)

$$R^2 = 0.9983$$

$$AC = \frac{RC}{X_i} = \frac{141.7667}{X_i} + 63.4776 - 12.96X_i + (0.9396)X_i^2$$

$$AC_i = 63.4776 - 12.9615X_i + 141.7667X_i + 0.9396X_i^2$$

$$MC = \frac{\partial TC}{\partial X_i} = 63.4776 - 2X(12.9615)X_i + 3 \times 0.9396X_i^2$$

If the cost curves are U-shaped Average Marginal cost curves then the theory suggests that the coefficient should satisfy following

1) β_1, β_2 and $\beta_4 > 0$
2) $\beta_3 < 0$
3) $\beta_3^2 < 3\beta_2\beta_4$

Check Your Progress 3

1) Define two significant methods to rectify the problem of multicollinearity?

.....

.....

.....

.....

.....

2) Describe the method of ridge regression.

.....

.....

.....

.....

.....

10.6 LET US SUM UP

This unit presents a clear understanding of the concept of multicollinearity in the regression model. The unit also presents a clear distinction of near and perfect multicollinearity. The unit familiarizes the consequences of presence of multicollinearity in regression model. The method of detection of multicollinearity has been highlighted in the unit. Finally various techniques that provide remedial measures including the concept of ridge regression have been explained in the unit.

10.7 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) The case of perfect multicollinearity mainly reflects the situation when the explanatory variables and perfectly correlated with each other implying the coefficient of correlation between the explanatory variables is 1.

2) This refers to the case when two or more explanatory variables are not exactly linear this reinforces the fact that collinearity can be high but not perfect. "High collinearity" refers to the case of "near" or imperfect" or high multicollinearity. Presence of multicollinearity implies "imperfect multicollinearity''

3) In the case of perfect multicollinearity it is not possible to obtain estimators for the parameters of the regression model. See Section 10.2 for details.

Check Your Progress 2

1) (i) In case of imperfect multicollinearity, some of the estimators are statistically not significant. But OLS estimates still retain their BLUE property that is, Best Linear Unbiased Estimators. Therefore, imperfect multicollinearity does not violate any of the assumptions, OLS estimators retain BLUE property. Being BLUE with minimum variance does not imply that the numerical value of variance will be small.

(ii) The R^2 value is very high but very few estimators are significant (t-ratios low). The example mentioned in earlier section where the demand function of good Y we computed using the earnings of individuals, reflects the situation where R^2 is quite high about 98% or 0.98 but only price variable slope coefficient has significant t-value. However, using F-test while testing overall significance $H_0 : R^2 = 0$, we reject the hypotheses that both prices and earnings have no effect on the demand of Y.

(iii) The ordinary least square OLS estimators mainly partial slope coefficients and their standard errors become very sensitive to small changes in the data, i.e. they then to be rentable. A small charge of data, the regression results change quite substantially as in case example of near or imperfect multicollinearity mentioned above, the standard errors go down and t-ratios have increased in absolute values.

(iv) Wrong signs of regression coefficients. It is a very prominent impact of presence of multicollinearity. In case of example where earnings of individuals were used in deriving demand curve of good Y, the earning

variable has the 'wrong' sign for the economic theory since the income effect usually positive unless it is case of inferior good.

2) Examining partial correlations: In case of three explanatory variables X_2, X_3 and X_4 very high or perfect multicollinearity between X_4 and X_2, X_3.

Subsidiary or auxiliary regressions: When one explanatory variables X is regressed on each of the remaining X variable and the corresponding R^2 is computed. Each of these regressions is referred as subsidiary or auxiliary regression. A regression Y on X_2, X_3, X_4, X_5, X_6 and X_7 with six explanatory variables. If R^2 comes out to be very high but few significant t-ratios or very few X coefficients are individually statistically significant then the purpose is to identify the source of the multicollinearity or existent of perfect or near perfect linear combination of other X^s.

For this we Regress X_2 on remaining X^s and obtain R_2^2 or also written as $R^2_{2.34567}$

Regress X_3 on remaining X^s, and obtain R_3^2 coefficient of determination also written as $R^2_{3.24567}$ each R_i^2 obtained will lie between 0 and 1. By testing the null hypothesis $H_0 : R_i^2 = 0$ by applying F-test. Let r_{23}, r_{24} and r_{34} represent pairwise correlation between X_2 and X_3, X_2 and X_4, X_3 and X_4 respectively suppose $r_{23} = 0.90$, reflecting high collinearity between X_2 and X_3. Considering partial correlations coefficient $r_{23.4}$ that indicators correlations coefficient between X_2 and X_3, Adding the influence of X_4 constant. If $r_{23.4} = 0.43$. Thus, partial correlation between X_2 and X_3 is low reflecting no high collinearity or low degree of collinearity. Therefore, pairwise correlation when replaced by partial correlation coefficients does not provide indicator of presence of multicollinearity.

3) Variance Inflation Factor (VIF): R^2 obtained variables auxiliary regression may not be completely reliable and is not reliable indicator of collinearity. In this method we modify the formula of var (b_2) and (b_3)

$$\text{var}(b_2) = \frac{\sigma^2}{\sum x_{2i}^2 \left(1 - R_2^2\right)}$$

$$= \frac{\sigma^2}{\sum X_{2i}^2} \cdot \left(\frac{1}{1 - R_2^2}\right)$$

$$\text{VIF} = \left(\frac{1}{1 - R_2^2}\right) \qquad \therefore V(b_2) = \frac{\sigma^2}{\sum x_{2i}^2} \text{V.I.F.}$$

Similarly, $V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}(VIF)$

VIF is variance inflation factor. As R^2 increases VIF $\frac{1}{1-R^2}$ increased thus inflating the variance and hence standard errors of b_2 and b_3

If $R^2 = 0$, $VIF = 1 \Rightarrow V(b_2) = \frac{\sigma^2}{\sum x_{2i}^2}$ and $V(b_3) = \frac{\sigma^2}{\sum x_{3i}^2}$

$\Rightarrow$ No collinearity

If $R^2 = 1$, $VIF = \infty \Rightarrow V(b_2) \to \infty, V(b_3) \to \infty$

If R^2 is high, however $var(b_2) \to \infty, var(b_3)$ does not only depend on R^2 (auxiliary coefficient of determination) or VIF. If also depends on σ^2 and $\sum x_{2i}^2$ it is possible that R_i^2 is high 0.91 but $var(b_2)$ could be lower due to low σ^2 or high $\sum x_{2i}^2$ thus $V(b_2)$be still lower resulting in high t value not showing any low t end thus defeating the indicator of multicollinearity. Thus R^2 obtained from and binary regression is only a surface indicator of multicollinearity.

Check Your Progress 3

1) (i) Dropping a variable from the Model: The simplest solution might seem to be to drop one or more of the collinear variables. However, dropping a variable from the model may lead to model specification error in either words, where we estimate the model without that variable, the estimated parameters of reduced model may turn out to be biased. Therefore, the best practical advice is not to drop or variable from an economically variable model first because the collinearity problem is serious. A variable which has t value of its coefficient greater than 1, then than variable should not be dipped as it will result in decrease in adjusted $\overline{R}^2$

(ii) Acquiring Additional Data or new sample: Acquiring additional data implies increasing the sample size can reduce the severity of collinearity problem.

$$V(b_2) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - R_2^2)}$$

Given σ^2 and R^2, if the sample size of X_3 increases $\Rightarrow \sum x_{3i}^2$ will increase as a result $V(b_3)$will tend to decrease and standard error b_3 will also.

2) In ridge regression we first standardise all the variables in the model. Go through Sub-Section 10.5.6 for details.

UNIT 11 HETEROSCDASTICTY*

Structure

11.0 OBJECTIVES

After going through this unit, you should be able to

- explain the concept of heteroscedasticity in a regression model;
- identify the consequences of heteroscedasticity in the regression model;
- explain the methods of detection of heteroscedasticity;
- describe the remedial measures for resolving heteroscedasticity;
- show how the use of deflators can help in overcoming the consequences of heteroscedasticity; and
- identify the correct functional form of regression model so that heteroscedasticity is avoided.

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

11.1 INTRODUCTION

A crucial assumption of the Classical Linear Regression Model (CLRM) is that the error term u_i in population regression function (PRF) is homoscedastic. It means that u_i has the same variance σ^2 throughout the population. An alternative scenario arises where the variance of u_i is σ_i^2. In other words, the error variance varies from one observation to another. Such cases are referred to as cases of heteroscedasticity.

11.2 HETEROSCEDASTICITY: DEFINITION

Let us first make a distinction between homoscedasticity and heteroscedasticity. This will help us in understanding the concept of heteroscedasticity better.

11.2.1 Homoscedasticity

Consider a 2-variable regression model, where the dependent variable *Y* is personal savings and the explanatory variable *X* is personal disposable income (or after-tax income).

As personal disposal income (PDI) increases, the mean or average level of savings also increases but the variances of savings around its mean value remains the same at all the levels of PDI. Such a case depicts the case of homoscedasticity or equal variance as shown in Fig. 11.1. In such cases, we have:

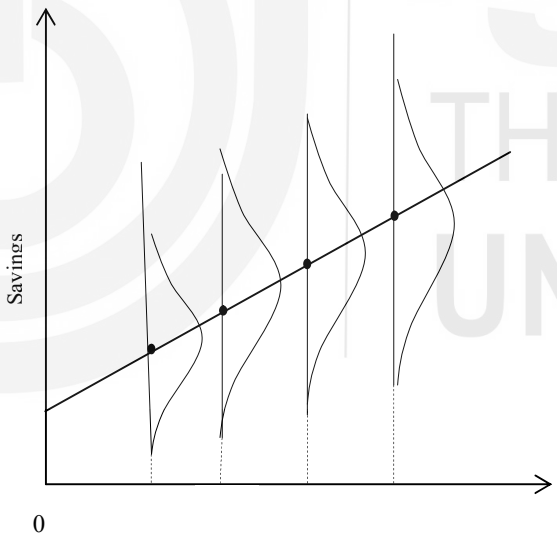


Fig.11.1: Case of Homoscedasticity

$$E(u_i^2) = \sigma^2 \quad \ldots (11.1)$$

We can alternatively express equation (11.1) as a case where:

$$V(u_i) = \sigma^2 \quad \ldots (11.2)$$

In Fig. 11.1, we see a case of homoscedasticity where the variance of the error term is a constant value, σ^2. This is expressed in the form of an equation as in

(11.2). Since the expected value of the error term is zero, the expression $V(u_i) = \sigma^2$ can also be written as $E(u_i^2) = \sigma^2$ as in equation (11.1).

11.2.2 Heteroscedasticity

As PDI increases, the average level of savings increases. However, the variance of savings does not remain the same at all the levels of PDI. This is the case of heteroscedasticity or unequal variance. In other words, high-income people, on average, save more than low-income people, but at the same time, there is more variability in their savings. This can be graphically represented as in Fig. 11.2. We now therefore have:

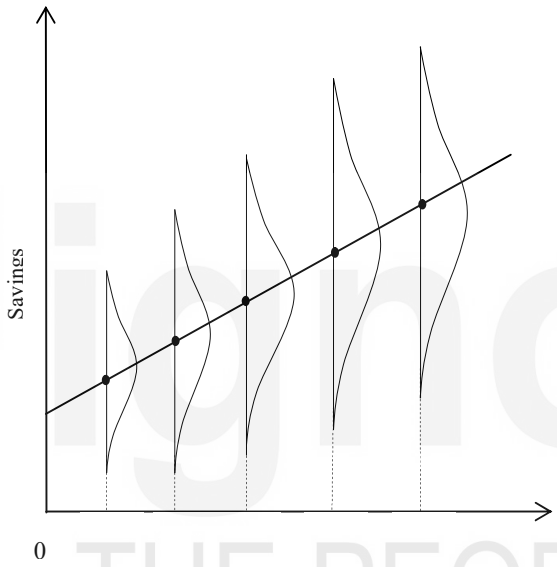


Fig. 11.2: Case of Heteroscedasticity

$$E(u_i^2) = \sigma_i^2 \quad or \quad V(u_i) = \sigma_i^2 \qquad \ldots (11.3)$$

The case of heteroscedasticity reflected in Fig.11.2 indicates that the error variance is not constant. It rather changes with every observation, like

$$V(u_i) = \sigma_i^2.$$

It is observed that heteroscedasticity is usually found in cross-sectional data and not so much in time series data. The reason for its occurrence more in cross-sectional data is mainly because, in the case of cross-sectional data, the members of population are like individuals, firms, industries, geographical division, state or countries. The data in such cases is collected at a point in time. Hence, the members of the population may be of different sizes: small, medium or large. This is referred to as the scale effect. In other words, due to what is called in economics as the 'scale effect', in cross sectional data we find cases of heteroscedasticity more commonly.

In the case of time series, on the other hand, the data of similar variables vary over a period of time. For instance, GDP (gross domestic product) or savings or unemployment varies over a period (like 1960 to 2008).

Check Your Progress 1

1) What is meant by heteroscedasticity?

.....

.....

.....

.....

.....

2) Is the problem of heteroscedasticity related to data? Comment.

.....

.....

.....

.....

.....

11.3 CONSEQUENCES OF HETEROSCEDASTICITY

To avoid the problem of heteroscedasticity, we have made one of the assumptions in the classical linear regression model that the error term is homoscedastic. However, in many regression models and actual data, the disturbance variance varies across observations. Consequently, the model suffers from specific impacts due to heteroscedastic error term.

The following are the characteristics of the OLS model in the presence of heteroscedasticity.

(i) The OLS estimators are linear function of the variables. The regression equation is also linear in its parameters.

(ii) The ordinary least squares (OLS) estimators are unbiased. This means the expected value of estimated parameters is equal to the true population parameters.

(iii) The OLS estimators though unbiased, are no longer with minimum variance, i.e., they are no longer efficient. In fact, even in large samples, the OLS estimators are not efficient. Therefore, the OLS estimators are not BLUE both in small as well as asymptotically large samples.

(iv) In light of the above, the usual formula for estimating variances of OLS estimator is biased, i.e., they are either upward biased (positive bias) or downward biased (negative bias). Note that when the OLS

overestimates the true variances of estimators, a positive bias is said to occur, and when it underestimates the true variances of estimators, we say that a negative bias occurs.

(v) The estimator of true population variance as given by $\hat{\sigma}^2 = \frac{\sum e_i^2}{df} = \frac{RSS}{df}$ is biased. That is

$$E(\hat{\sigma}^2) \neq \sigma^2 \qquad \ldots (11.4)$$

We know that the degrees of freedom for testing an estimated parameter is (n – k), where k is the number of parameters (or explanatory variables) in the regression model. For example, if there are three explanatory variables, d.f. = (n – 3). In the two variables case, df = (n – 2). Note that we are counting the intercept estimate for this purpose of determining the d.f.

(vi) Equation (11.4) implies that in the presence of heteroscedasticity, the estimated value of error variance is not equal to the true population error variance. In view of this, the usual confidence interval and hypothesis testing based on t and F distributions are unreliable (since, the estimator of the error variance is biased). Therefore, the possibility of making wrong inferences (Type–II error) is very high. As a result, in the presence of heteroscedasticity, the results of the usual hypothesis-testing are not reliable raising the possibility of drawing misleading conclusions.

Check Your Progress 2

1) State any two important consequences of heteroscedasticity.

.....

.....

.....

.....

.....

2) In the presence of heteroscedasticity, the OLS estimator will either overestimate or underestimate the error variance. Justify the statement.

.....

.....

.....

.....

.....

11.4 DETECTION OF HETEROSCEDASTICITY

So far, we have discussed the consequences of heteroscedasticity. Now let us discuss how heteroscedasticity can be detected. There are quite a few methods of detecting heteroscedasticity. Some of these methods are described below.

11.4.1 Graphical Examination of the Residuals

We can begin with examining the residuals obtained from the fitted regression line. The residual plot of squared residuals is an indicator of the existence of heteroscedasticity. Since the error terms u_i are not observable, we examine the residuals, e_i.

A plot of the residuals can give us various types of diagrams as in Fig. 11.3.

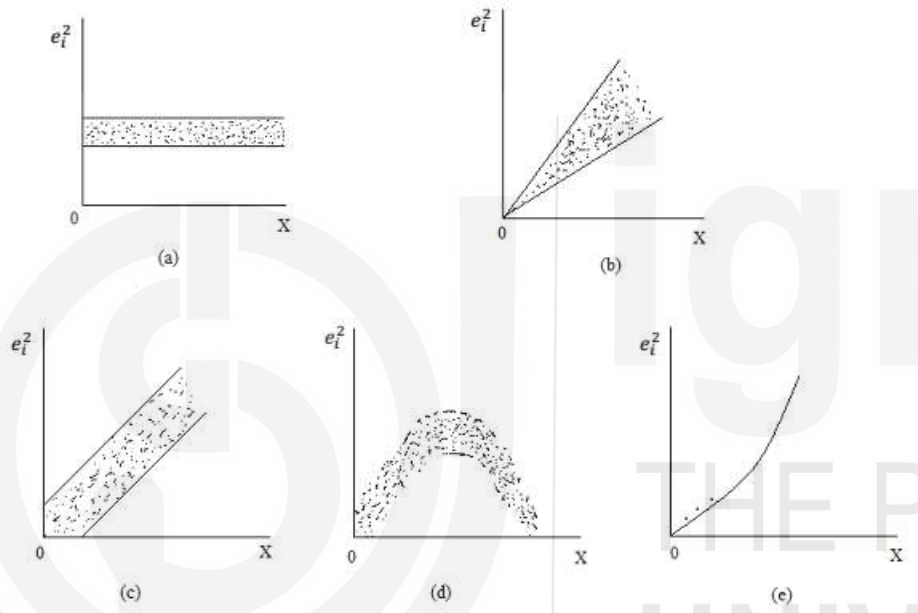


Fig. 11.3: Cases of Homoscedasticity and Heteroscedasticity

In the five situations depicted in Fig. 11.3, we see that Case (a) represents homoscedasticity, i.e., $V(u_i) = \sigma^2$ whereas in the remaining four cases viz., (b), (c), (d) and (e) represent heteroscedasticity, i.e., $V(u_i) = \sigma_i^2$.

11.4.2 Park-Test

If there is heteroscedasticity in a data set, the heteroscedastic variance σ_i^2 may be systematically related to one or more explanatory variables. Therefore, we can regress σ_i^2 on one or more explanatory variables such as

$\sigma_i^2 = f(X_i)$

$\ln\sigma_i^2 = \beta_1 + \beta_2 \ln X_i + v_i$ … (11.5)

In equation (11.5), a non-linear (double-log) regression is run to establish a relationship between the error variance and the explanatory variable with v_i

taken as the residual term. When σ_i^2 are not known, we take the residual term e_i as proxies for u_i. Therefore, we have

$$\ln e_i^2 = \beta_1 + \beta_2 \ln X_i + v_i \quad \ldots (11.6)$$

Now, Park test for detecting heteroscedasticity involves the following steps:

a) Run the original regression in equation (11.5) despite the heteroscedasticity problem.

b) From the regression obtain e_i and square them. Then take the logs of e_i^2.

c) Run the double-log form regression as indicated in equation (11.6) using an explanatory variable in the original model (in the case of more than one explanatory variable). Then run the regression against each *X* variable. In other words, we run the regression against $\hat{Y}_i$, the estimated value of Y_i.

d) Test the null hypothesis $\beta_2 = 0$, i.e., there is no heteroscedasticity.

e) A statistically significant relationship implies that the null hypothesis of no heteroscedasticity is rejected. It suggests the presence of heteroscedasticity which requires remedial measures.

f) If the null hypothesis is not rejected, then it means we accept $\beta_2 = 0$ and the value of β_1, that is, the value of the intercept can be accepted as the common, homoscedastic variance σ^2.

11.4.3 Glejser Test

The Glejser Test is similar to the Park Test. The steps to carry out the Glejser test are as follows:

a) Obtain the residual e_i from the original model.

b) Take absolute value $|e_i|$ of the residuals

c) Regress the absolute values of $|e_i|$ on the *X* variable that is expected to be closely associated with heteroscedastic variance σ_i^2.

d) You can take various functional forms of X_i. Some of the functional forms suggested by Glejser are

$$|e_i| = \beta_1 + \beta_2 X_i + v_i \quad \ldots (11.7)$$

$$|e_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i \quad \ldots (11.8)$$

$$|e_i| = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + v_i \quad \ldots (11.9)$$

The above means that the Glejser test suggests various plausible (linear as well as non-linear) relationships between the residual term and the explanatory variable to investigate the presence of heteroscedasticity.

e) For each of the cases given, test the null hypothesis that there is no heteroscedasticity, i.e., $H_0: \beta_2 = 0$ (no heteroscedasticity).

f) If H_0 is rejected we conclude that there is evidence of heteroscedasticity.

You should note that the error term v_i can itself be heteroscedastic as well as serially correlated. Thus, in the case of Glesjer test also, we follow the same steps as in the Park Test. The difference between the two tests is in the functional forms to be considered.

11.4.4 White's General Test

Let us consider the following PRF:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \ldots (11.10)$$

The steps to carry out White's general test for heteroscedasticity are as follows:

a) Estimate the population regression equation (11.10) by OLS and obtain the residuals e_i.

b) Find the square of the residuals e_i^2.

c) Run the following auxiliary regression:

$$e_i^2 = A_1 + A_2 X_{2i} + A_3 X_{3i} + A_4 X_{2i}^2 + A_5 X_{3i}^2 + A_6 X_{2i} X_{3i} + v_i \ldots (11.11)$$

d) Obtain the coefficient of determination R^2 from the auxiliary regression under the null hypothesis that there is no heteroscedasticity (i.e., all the slope coefficient are zero). That is,

$$H_0: A_2 = A_3 \ldots A_6 = 0 \quad \ldots (11.12)$$

The null hypothesis given at equation (11.12) implies that all the partial slope coefficients are simultaneously zero. Note that we do not include the intercept term A_1 in equation (11.12).

e) Test the null hypothesis in equation (11.12) by using the chi-square distribution as follows:

$$nR^2 \sim \chi_{k-1}^2 \quad \ldots (11.13)$$

Equation (11.13) tells us that the product of sample size (n) and the coefficient of determination (R^2) follows χ^2 distribution with degrees of freedom (k–1). Here k is the number of regressors in the auxiliary regression (equation 11.11).

f) If $\chi_{calculated}^2 > \chi_{critical}^2$ we reject the H_0, and conclude that the null hypothesis of homoscedasticity is to be rejected, i.e., there is heteroscedasticity. Alternatively, we can also decide on the basis of the p value (readily given by econometric softwares). If the p value is < 0.05, we reject H_0. If $\chi_{calculated}^2 < \chi_{critical}^2$. On the other hand, if $p > 0.05$ we do not reject the null hypothesis of no heteroscedasticity. This implies the existence of homoscedasticity.

11.4.5 Goldfeld-Quandt Test

The Goldfeld-Quandt (G-Q) test is applicable if heteroscedasticity is related to only one of the explanatory variables. Let us assume that the error variance σ_i^2 is related to one of the explanatory variables (say, X_i) in the regression model.

Suppose σ_i^2 is positively related to X_i as given below.

$$\sigma_i^2 = \sigma^2 X_i^2 \qquad \ldots (11.14)$$

In order to carry out the G-Q test we proceed as follows:

a) Arrange the observations in increasing order of X_i

b) Omit some of the observations (say, C out of the no observations in the sample) in the middle of the series. There is no hard and fast rule for the exact value of C and the choice is quite arbitrary. In practice about one fourth observations are omitted.

c) Run a regression on the first $n_1 = (n - C)/2$ observations. Find out the error sum of squares for this regression, i.e., ESS_1.

d) Run a regression on the last $n_2 = (n - C)/2$ observations. Find out the error sum of squares for this regression, i.e., ESS_2.

e) Take the following null hypothesis:

$$H_0: \sigma_i^2 = \sigma^2 \qquad \ldots (11.15)$$

f) Find out the ratio:

$$\lambda = \frac{RSS_1 / \frac{n_1 - C - 2k}{2}}{RSS_2 / \frac{n_2 - C - 2k}{2}}$$

In case $n_1 = n_2$, the above ratio becomes

$$\lambda = \frac{RSS_1}{RSS_2} \qquad \ldots (11.6)$$

The above ratio (λ) follows F-distribution with degrees of freedom

$$\left(\frac{n_1 - C - 2k}{2}, \quad \frac{n_2 - C - 2k}{2} \right) \qquad \ldots (11.17)$$

g) We compare the value of λ obtained above with the tabulated value of F given at the end of the book. If $\lambda > F_{critical}$ we reject $H_0: \sigma_i^2 = \sigma^2$and conclude that there is heteroscedasticity in error variance. It implies $\sigma_i^2 \neq \sigma^2$. If $\lambda < F_{critical}$ we do not reject H_0. We conclude that there is homoscedasticity in error variance, i.e., $\sigma_i^2 = \sigma^2$.

Check Your Progress 3

1) State the steps in conducting the Park test for detection of heteroscedasticity.

.....

.....

.....

.....

.....

11.5 REMEDIAL MEASURES OF HETEROSCEDASTICITY

Heteroscedasticity means that the OLS estimators are unbiased but no longer efficient; not even in large samples. Therefore, if heteroscedasticity is present, it is important to seek remedial measures. For proceeding with remedial measures, it is important to know if the true error variance σ_i^2 is known or not. In such cases, use of a 'deflator' may help rectify the problem of heteroscedasticity. We will learn about the use of deflators in this section.

11.5.1 Case I: σ_i^2 is Known

If we know σ_i^2, we can use the method of Weighted Least Squares (WLS). We explain the procedure of carrying out WLS below.

Let us consider the two-variable Population Regression Function (PRF).

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \ldots(11.18)$$

Let us assume that u_i has heteroscedastic error variance. Here, since the true variance is known, we can use it to divide the equation (11.18) by σ_i. By dividing both sides of (11.18) by σ_i, we obtain:

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{1}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \frac{u_i}{\sigma_i} \qquad \ldots (11.19)$$

Note that the error term gets transformed due to the division by σ_i. Let the new error term be v_i. Squaring the new error term we get:

$$v_i^2 = \frac{u_i^2}{\sigma_i^2} \qquad \ldots(11.20)$$

Since the variance of error term is given by $var(v_i) = E(v_i^2)$, taking the expectation of both sides of the equation (11.20) we get:

$$E(v_i^2) = E\left(\frac{u_i^2}{\sigma_i^2}\right)$$

$$= \left(\frac{1}{\sigma_i^2}\right).E(u_i^2)$$

$$= \frac{\sigma_i^2}{\sigma_i^2} = 1$$

Thus, the transformed error-term v_i is homoscedastic. Therefore, equation (11.19) can be estimated by the usual OLS method. The OLS estimators of β_1 and β_2 thus obtained are called the Weighted Least Squares (WLS) estimators.

11.5.2 Case II: σ_i^2 is Unknown

When the error variance σ_i^2 is not known, we need to make further assumptions to use the WLS method. Here, we consider the following two cases.

(i) **Error variance σ_i^2 is Proportional to X_i**

In this case, we follow what is called as the square root transformation. The proportionality assumption means that:

$E(u_i^2) = \sigma^2 X_i$

Or, $V(u_i) = \sigma^2 X_i$ … (11.21)

Now, the square root transformation requires that we divide both sides of equation (11.18) by $\frac{1}{\sqrt{X_i}}$ to get:

$$\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \frac{X_i}{\sqrt{X_i}} + \frac{u_i}{\sqrt{X_i}}$$

$$= \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + v_i \quad \text{… (11.22)}$$

where $v_i = \frac{u_i}{\sqrt{X_i}}$ … (11.23)

The error term in equation (11.23) is a transformed error term. In order to see whether v_i is devoid of heteroscedasticity, we square both the sides of equation (11.23) to get:

$v_i^2 = \frac{u_i^2}{X_i}$ … (11.24)

Now, the variance of the transformed error term, i.e., equation (11.24) is:

$E(v_i^2) = \frac{E(u_i^2)}{X_i} = \frac{\sigma^2 X_i}{X_i}$ ... (11.25)

$= \sigma^2 \Rightarrow$ homoscedasticity

Thus, when we apply the square root transformation ($v_i = \frac{u_i}{\sqrt{X_i}}$), we could make the error variance to become homoscedastic.

(ii) Error Variance is Proportional to X_i^2

Here, we have:

$E(u_i^2) = \sigma X_i^2$ … (11.27)

$$V(u_i) = \sigma X_i^2$$

Dividing both sides of equation (11.18) by X_i,

$$\frac{Y_i}{X_i} = \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + \left(\frac{u_i}{X_i}\right)$$

$$= \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + v_i \quad \ldots (11.28)$$

Equation (11.28) is the transformed PRF in which the error term is:

$$v_i = \frac{u_i}{X_i}, \quad \ldots (11.29)$$

Squaring both the sides of equation (11.29), we get:

$$v_i^2 = \frac{u_i^2}{X_i^2} \quad \ldots (11.30)$$

The variance of the error term of the transformed equation in (11.30) is homoscedastic because:

$$E(v_i^2) = \frac{E(u_i^2)}{X_i^2} = \frac{\sigma X_i^2}{X_i^2} = \sigma \quad \ldots (11.31)$$

11.5.3 Re-Specification of the Model

Instead of speculating about σ_i^2, sometimes choosing a different functional form can reduce heteroscedasticity. For instance, instead of running the usual regression model, we can estimate the model in its log form.

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad \ldots (11.32)$$

In many cases transforming original model as above will take care of the problem of heteroscedasticity.

We used the word 'deflator' in the beginning of this section. The cases we have considered above basically involve dividing both sides of the original regression model by a known value to transform the variables. Such transformation of variables by division amounts to deflating the original values. The known values used to perform the division act are known as the 'deflators'.

Check Your Progress 4

1) How does the use of deflators work as a solution for the problem of heteroscedasticity?

.....

.....

.....

.....

2) Explain how the usage of deflators serve to tackle the problem of heteroscedasticity when the error variance is proportional to X_i^2.

.....

.....

.....

.....

.....

11.6 LINEAR VERSUS LOG – LINEAR FORMS

The regression model can be run in various functional forms depending upon: (i) the relationship of dependent and independent variable, and (ii) the data. Suppose there is a choice of running two types of regression models: (i) a linear regression model, and (ii) a log-linear model. To help decide in such cases, a test for the selection of the appropriate functional form for regression is proposed by Mackinnon, White and Davidson (MWD). The MWD test is applied as follows:

Let there be two distinct functional forms of a regression like:

Model 1: $Y_i = \beta_1 + \beta_2 X_i + u_i$ …….(11.33)

Model 2: $\ln Y_i = \beta_1 + \beta_2 lnX_i + u_i$ …….(11.34)

In Model 1, the dependent variable is linearly related to one (or more than one) of the *X*s. In Model 2, the relationship between the dependent and independent variable is non-linear. The MWD test involves considering a null and an alternate hypothesis as follows:

H_0: Linear Model, i.e., *Y* is a linear function of regressors (equation (11.33))

H_1: Log- Linear Model, i.e., ln *Y* is a linear function of lnX_i (equation (11.34))

Following are the steps for carrying out the MWD test:

(i) Estimate the linear model and obtain the estimated *Y* values. Let the estimated *Y* values be denoted as Y_f.

(ii) Estimate the log-linear model and obtain the estimated ln*Y* values. Let the estimated values of the log-linear *Y* be denoted as $\ln Y_f$.

(iii) Obtain $Z_1 = (\ln Y_f - Y_f)$

(iv) Regress *Y* on X_s and Z_1 obtained in Step (iii) Reject H_0 if the coefficient of Z_1 is statistically significant by the usual *t*-test.

(v) Obtain $Z_2 = (\text{antilog } \ln Y_f - Y_f)$

(vi) Regress log of Y on the logs of X_s and Z_2. Reject H_1 if the coefficient of Z_2 is statistically significant by the usual t-test.

Suppose the linear model I in equation (11.33) is in fact the correct model. In that case, the constructed variable Z_1 should not be statistically significant in Step (iv). For, in that case the estimated Y values from the linear model and those estimated from the log-linear model (after taking their antilog values for comparative purposes) in equation (11.34) should not be different. The same logic applies to the alternative hypothesis H_1.

Check Your Progress 5

1) Outline the MWD test for choosing the appropriate functional form of the regression model between its linear and log-linear forms.

.....

.....

.....

.....

.....

11.7 LET US SUM UP

In this Unit, we have discussed the concept of heteroscedasticity in regression models. The unit outlines the consequences of the presence of heteroscedasticity and the methods of its detection. Various techniques to provide remedial measures are explained in the unit. The remedial measures involve understanding of the use of deflators. The unit has also explained a method for the choice of selecting the functional form by way of the MWD test.

11.8 ANSWERS/ HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) A crucial assumption of the Classical Linear Regression Model CLRM is that the error term u_i is population regression function (PRF) is homoscedastic, i.e., they have the same variance σ^2. However, if the variance of u_i is σ_i^2 (in other words, it varies from one observation to another), then the situation is referred to as heteroscedasticity.

2) Heteroscedasticity is usually found is cross-sectional data and not in time series data. This is because, in the case of cross-sectional data, the members of population are in the form of individual firms, industries, geographical division, state or countries. The data collected for such units at a point of time from the members of population may be of different sizes: small, medium or large firms. This is referred to as scale effect.

Due to the scale effect, in cross-sectional data, there is a greater chance of coming across heteroscedasticity in the error terms.

Check Your Progress 2

1) The OLS estimators are unbiased but they no longer have minimum variance, i.e., they are no longer efficient. Even in large samples the OLS estimators are not efficient. Therefore, the OLS estimators are not BLUE in small as well as large samples (asymptotically).

   The usual formula for estimating the variances of OLS estimator are biased i.e. there is either upward bias (positive bias) or downward bias (negative bias).

2) The OLS estimator of error variance is a biased estimator. Thus it will either overestimate or underestimate. In fact, the OLS estimator of error variance is inefficient, thereby meaning that it is very high; thus it is always an overestimate.

Check Your Progress 3

1) In the presence of heteroscedasticity, the heteroscedastic variance σ_i^2 may be systematically related to one or more explanatory variables. Therefore, we can regress σ_i^2 on one or more of *X*- variables as:

   $\sigma_i^2 = f(X_i)$ or $\ln\sigma_i^2 = \beta_1 + \beta_2 \ln X_i + v_i$

   where $v_i = new$ residual term. If σ_i^2 are not known, estimated e_i can be used as proxies for u_i. A statistically significant relationship implies that the null hypothesis of no heteroscedasticity is rejected suggesting the presence of heteroscedasticity which requires remedial measures. If null hypothesis is not rejected then it means we accept $\beta_2 = 0$ and value of β_1 can be taken as the common, homoscedastic variance σ^2.

2) Heteroscedasticity means that the OLS estimators are unbiased but estimators are no longer efficient, not even in large samples. This lack of efficiency makes the conventional hypothesis testing of OLS estimators unreliable. For remedial measures, it is important to know whether the true error variance σ_i^2 is known or not. In such cases, use of deflators will help rectify the problem of heteroscedasticity. Various deflators can be used to convert the error variance ti make them homoscedastic.

   When σ_i^2 is known, the method of Weighted Least Squares (WLS) can be considered. In this, the error variance σ_i^2 is used to divide both sides of the equation by σ_i. See Section 11.5 for details.

3) The estimated residuals show a pattern similar to earlier case I, but error variance is not linearly related to *X* but increases proportional to square of X. Hence, $E(u_i^2) = \sigma X_i^2$ and $V(u_i) = \sigma X_i^2$. Dividing both sides by X_i, we get:

   $$\frac{Y_i}{X_i} = \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + \left(\frac{u_i}{X_i}\right)$$

   $$= \beta_1\left(\frac{1}{X_i}\right) + \beta_2 + v_i$$

$v_i = \frac{u_i}{X_i}, v_i^2 = \frac{u_i^2}{X_i^2}$

$E(v_i^2) = \frac{E(u_i^2)}{X_i^2} = \frac{\sigma X_i^2}{X_i^2} = \sigma$

Thus, the transformed equation is homoscedastic.

Check Your Progress 5

1) The test for selection of the appropriate functional form for regression as proposed by Mackinnon, White and Davidson is known as MWD Test. The MWD test is used to choose between the two models. See Section 11.6 for details.

UNIT 12 AUTOCORRELATION *

Structure

12.0 OBJECTIVES

After going through this unit, you should be able to:

- outline the concept of autocorrelation in a regression model;
- describe the consequences of presence of autocorrelation in the regression model;
- explain the methods of detection of autocorrelation;
- discuss the procedure of carrying out the Durbin-Watson test for detection of autocorrelation;
- elucidate the remedial measures for resolving autocorrelation; and
- outline the procedure of dealing with situations where autocorrelation exists in models with a lagged dependent variable.

12.1 INTRODUCTION

In the previous unit, you studied about heteroscedasticity. You saw that heteroscedasticity is a violation of one of the assumptions of the Classical Linear Regression Model (CLRM), viz., homoscedasticity. If the variance of the error term is not constant across all observations, then we have the problem of heteroscedasticity. In this unit, we discuss about the violation of another assumption of the CLRM. Recall that one of the assumptions about the error

* Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

terms is that the error term of one observation is not correlated with the error term of another observation. If they are correlated, then the situation is said to be one of autocorrelation. This is also called as the problem of serial correlation. This can be present in both cross-section as well as time series data. Let us discuss the concept of autocorrelation in a little more detail.

12.2 CONCEPT OF AUTOCORRELATION

The classical linear regression model (CLRM) assumes that the correlation among various error terms is zero. We know that heteroscedasticity is associated more with cross sectional data. Autocorrelation is usually more associated with time series data. Of course, autocorrelation can be present even in cross-section data. Some authors use the term autocorrelation only for time-series data. They use the term 'serial correlation' for describing autocorrelation in cross-section data. Many authors use the terms autocorrelation and serial correlation as synonyms. They use the term across both cross-section as well as time-series data.

Autocorrelation occurring in cross-sectional data is also sometimes called spatial correlation (correlation in space rather than in time). In CLRM we assume that there is no autocorrelation. This implies:

$$E(u_i, u_j) = 0 \qquad i \neq j \qquad \ldots(12.1)$$

Equation (12.1) means that the stochastic error term associated with one observation is not related to or influenced by the disturbance term associated with any other observation. For instance, the labour strike in one quarter affecting output may not affect the output in the next quarter. This implies there is no autocorrelation in the time series. Similarly, in a cross-section data of family consumption expenditure, the increase in one family's income on consumption expenditure in not expected to affect the consumption expenditure of another family. In the example of output affected due to labour strike above, if $E(u_i, u_j) \neq 0,\ i \neq j$, this implies a situation of autocorrelation. This means the disruption caused by the strike in one quarter is affecting the output in the next quarter. Similarly, increase in consumption expenditure of one family may influence the consumption expenditure of other families in the neighbourhood due to the 'demonstration effect' (cross-sectional data). It is thus more a case of spatial correlation. It is therefore important to analyse the data carefully to bring out what exactly is causing the correlation among the disturbance terms. Let us see more carefully the different situations or cases of autocorrelation as depicted in Fig.12.1. In panels (a) to (d) of Fig. 12.1 we find distinct pattern among u_t . In panel (e) of Fig. 12.1 we do not see any such pattern. Note that since autocorrelation is seen mostly in time series data, we use the subscript 't' in place of 'i' to indicate individual observations. Let us now study the reasons of autocorrelation with some specific examples from economics.

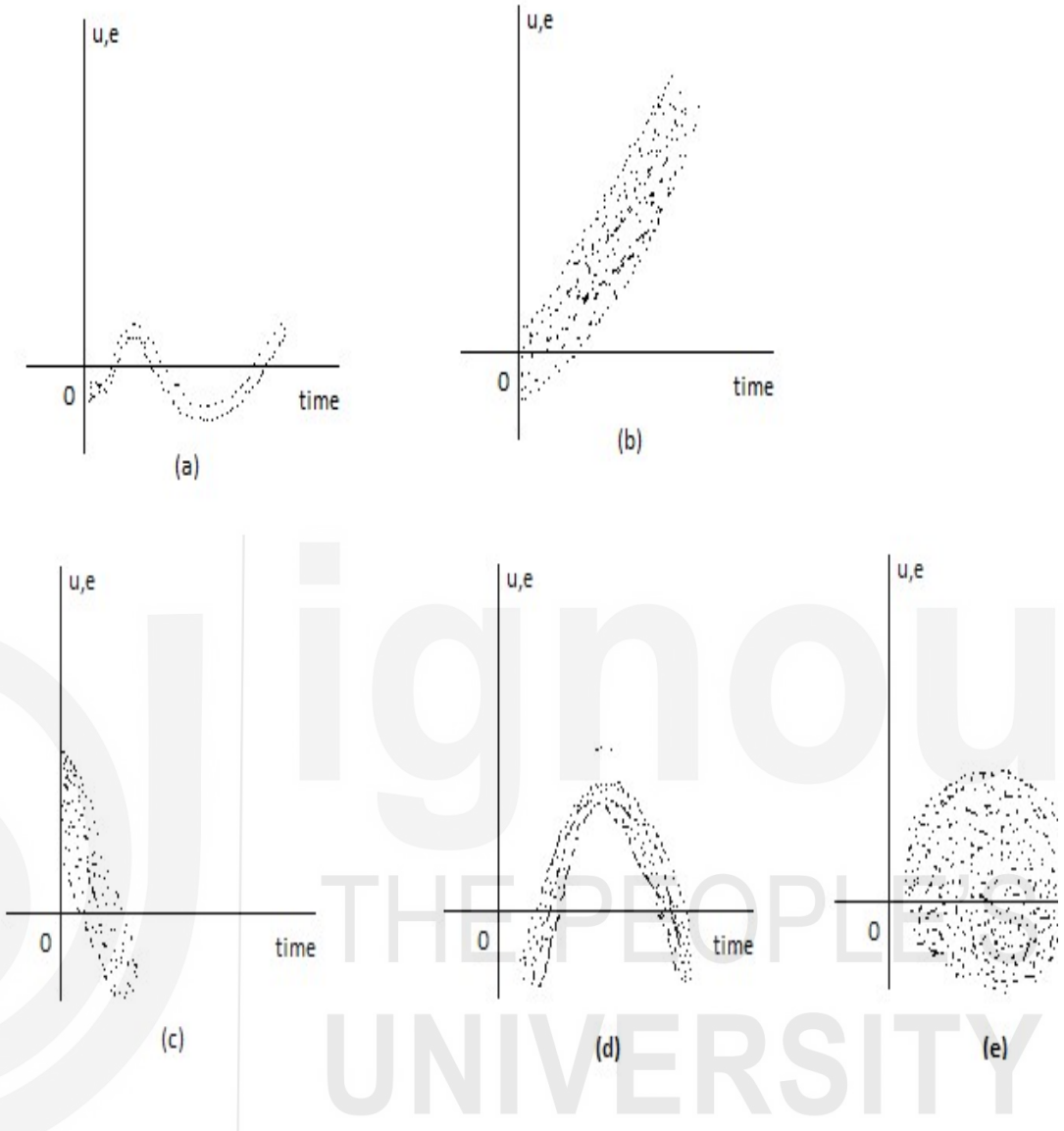


Fig. 12.1: Cases of Autocorrelation

12.3 REASONS FOR AUTOCORRELATION

The various reasons for the presence of autocorrelation can be discussed under the following broad heads.

(a) Inertia or Sluggishness

Most of the economic time series data displays inertia or sluggishness. For instance, gross domestic product (GDP), production, employment, money supply, etc. reflect recurring and self-sustaining fluctuations in economic activity. When an economy is recovering from recession, most of the time series will be moving upwards. This means any subsequent value of a series at one point of time is always greater than its previous time value.

Such a momentum continuous till it slows down due to, say, a factor like increase in taxes or interest or both. Hence, in regressions involving time series data, successive observations would generally be inter-dependent or correlated. Such an uptick effect is termed as 'inertia' which literally means a situation that continues to hold in a similar manner for many successive time periods. We see its opposite effect in periods of recession when most of the economic activity will be suffering, i.e., will be sluggish.

(b) Specification Error in the Model

By an incorrect specification of model, certain important variables that should be included in the model may not be included (i.e. a case of under-specification). If such model-misspecification occurs, the residuals from such an incorrect model will exhibit systematic pattern. If the residuals show a distinct pattern, it gives rise to serial correlation.

(c) The Cobweb Phenomenon

Many agricultural commodities reflect what is called as a 'cobweb phenomenon'. In this, supply reacts to price with a lag of time. This is mainly because supply decisions take time to implement. In other words, there is a gestation period involved. For instance, farmers' decision to plant crop might depend on the prices prevailing in the previous year's supply position or function. This can be written as:

$$S_t = \beta_1 + \beta_2 P_{t-1} + u_t \qquad \ldots (12.2)$$

In (12.2), the error term u_t may not be purely random. This is because, if the farmers over-produce in year *t*, they are likely to under-produce in year (t + 1) since they want to clear away the unsold stock. This usually leads to a cobweb pattern.

(d) Data Smoothing

Sometimes we need to average the data presented. Considering averages implies 'data smoothing'(see Unit 5 of BECC 109 for an example). We may prefer to convert monthly data into quarterly data by averaging the data over every three months. However, this smoothness, desired in many contexts, may itself lead to a systematic pattern in disturbances, resulting in autocorrelation.

Autocorrelation may be positive or negative depending on the data. Generally, economic data exhibits positive autocorrelation. This is because most of them either move upwards or downwards over time. Such a trend continues at least for some time i.e. some months, or quarters. This means, they are not generally expected to exhibit a sudden upward or downward movement unless there is a reason or a shock.

12.4 CONSEQUENCES OF AUTOCORRELATION

When the assumption of no-autocorrelation is violated, the estimators of the regression model based on sample data suffers from certain consequences. More specifically, the OLS estimators will suffer from the following consequences.

a) The least squares estimators are still linear and unbiased. In other words, the estimated values of parameters continue to be unbiased. However, they are not efficient because they do not have minimum variance. Therefore, the usual OLS estimators are not BLUE (best linear unbiased estimators).

b) The estimated variances of OLS estimators $(b_1 \text{ and } b_2)$ are biased. Hence, the usual formula used to estimate the variances, and their standard errors underestimate the true variances and standard errors. Consequently, the decision of rejecting a parameter on the basis of *t*-values, concluding that a particular coefficient is statistically different from zero, would be an incorrect conclusion. In other words, the usual *t* and *F* tests become unreliable.

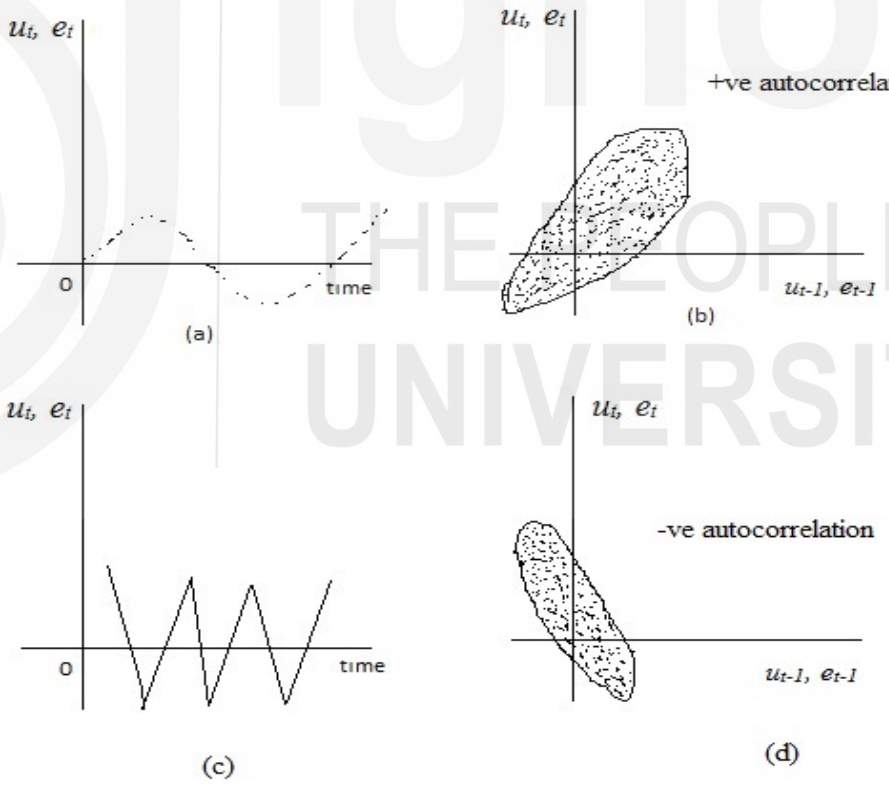


Fig. 12.2: Patterns of the Error Term in Autocorrelation

a) As a direct consequence of the above, the usual formula for estimating the population error variance, viz., $\hat{\sigma}^2 = (\text{RSS}/\text{df})$ yields a biased estimator

of true σ^2. In particular, it underestimates the true σ^2. As a consequence, the computed R^2 becomes an unreliable measure of true R^2.

Fig. 12.2 shows the pattern of error terms under different situations of autocorrelation. Note that since the population error terms (u_t) are not known, we are plotting the sample residuals (e_t).

Check Your Progress 1 [Answer the questions in 50-100 words within the space given]

1) What is meant by autocorrelation in a regression model?

……

……

……

……

……

2) In which type of data the problem of autocorrelation is more common? Why?

……

……

……

……

……

3) State the broad reasons for autocorrelation.

……

……

……

……

……

4) Enumerate the consequences of autocorrelation.

……

……

……

……

……

12.5 DETECTION OF AUTOCORRELATION

There are many methods of detecting the presence of autocorrelation. Let us discuss them now.

12.5.1 Graphical Method

A visual examination of OLS residuals e_t quite often conveys the presence of autocorrelation among the error terms u_t. Such a graphical presentation (Fig. 12.3) is known as the 'time sequence plot'. The first part of this figure does not show any clear pattern in the movement of the error terms. This means there is an absence of autocorrelation. In the lower part of Fig. 12.3, you will notice that the correlation between the two residual terms is first negative and then becomes positive. Therefore, plotting the sample residuals gives us the first indication on the presence or absence of autocorrelation.

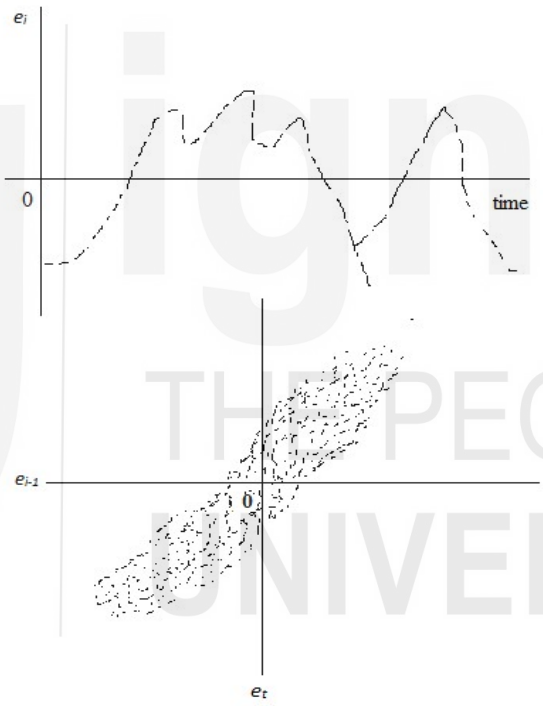


Fig. 12.3: Graphical Method for Detection of Autocorrelation

12.5.2 Durbin-Watson Test

The Durbin-Watson test, or the DW test as it is popularly called, is an analytical method of detecting the presence of autocorrelation. Its statistic is given by:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad \ldots (12.3)$$

Equation (12.3) defines the *d*-statistic suggested by Durbin-Watson as the ratio of the sum of squared differences in the successive residuals to the residual sum of squares. For computing the *d*-statistic, we take the sample size to be (*n*–1) since

one observation is lost in taking the successive differences. There are certain assumptions underlying the *d*-statistic. These are:

(a) The regression model includes an intercept term. Therefore, this method cannot be used to determine autocorrelation in regression models without the intercept term (i.e. regression equation which passes through the origin).

(b) The *X* variables are non-stochastic, i.e., their values are fixed in repeated samples.

(c) The error term evolves as follows :

$$u_t = \rho u_{t-1} + v_t, \qquad -1 \leq \rho \leq 1 \qquad \ldots (12.4)$$

Equation (12.4) states that the value of error term at time period *t* is dependent on the value of the error term in time-period (*t*–1) and a purely random term v_t. The extent of dependence on past value is measured by ρ which lies between –1 and 1.

The regression model given in equation (12.4) is referred to as the first-order auto-regression scheme. It is denoted by *AR(1)*. The usage of the term 'autoregressive' implies that the error term u_t is regressed on its own lagged value of one period, i.e., u_{t-1}. It is therefore called the first-order autoregressive scheme. If we include 2 lagged values (i.e., u_{t-1} and u_{t-2}) then we have the *AR(2)* scheme. Likewise, when we extend the number of lagged values to '*p*', we have the *AR(p)* scheme.

(d) The regression model does not contain any lagged value of the dependent variable as one of the explanatory variables. In other words, the test is not applicable to models like:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t \qquad \ldots (12.5)$$

where Y_{t-1} is the one-period lagged-value of the dependent variable *Y*. Models of the above type are known as auto-regressive (AR) models. For such cases, the *d*-statistic cannot be used.

We can estimate ρ from equation (12.4) as follows:

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}$$

[Recall that the estimator of b_2 in the two variable regression model is $b_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$.

We apply the same logic to derive $\hat{\rho}$ above]

We can expand equation (12.3) to obtain

$$d = \frac{\sum e_t^2 + \sum e_{t-1}^2 - 2\sum e_t e_{t-1}}{\sum e_t^2}$$

The above can be approximated to

$$d \approx 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}\right)$$

We can take an approximate value of d as:

$$d \approx 2(1 - \hat{\rho}) \quad \ldots (12.6)$$

where the symbol $\approx$ denotes 'approximately'. In equation (12.6), $\hat{\rho}$ is an estimator of the first order autocorrelation scheme. Table 12.1 presents the value of the d-statistic for different values of $\hat{\rho}$.

From Table 12.1 we find that $0 \leq d \leq 4$. The Durbin-Watson statistic thus provides a lower limit d_L and an upper limit d_U. The computed value of d is therefore a value between 0 and 4. From such a value, we can infer on the nature of autocorrelation as follows:

a) If d is closer to 0, there is evidence of positive autocorrelation.

b) If d is closer to 2, there is evidence of no autocorrelation.

c) If d is closer to 4, there is evidence of negative autocorrelation.

Table 12.1: Value of *d*-Statistic according to $\hat{\rho}$

Value of $\hat{\rho}$	Implication	Value of *d*-statistic
$\hat{\rho} = -1$	Perfect negative autocorrelation	4
$\hat{\rho} = 0$	No autocorrelation	2
$\hat{\rho} = 1$	Perfect positive autocorrelation	0

The steps in applying the DW test are therefore the following:

1. Run the OLS regression and obtain the residuals e_t.
2. Compute d as:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

3. Find out the critical Table values d_L and d_U for given sample size and given number of explanatory variables.

Follow the decision rule, as depicted in Fig. 12.4.

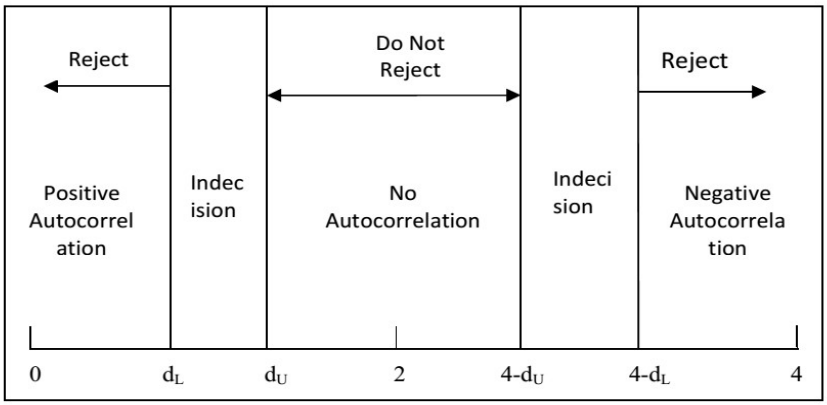


Fig. 12.4: Range of Values of Durbin-Watson Statistic

One drawback of the *d*-test is that it has two zones of indecision viz. $d_L < d < d_U$ and ($4 - d_U < d < 4 - d_L$).

12.5.3 The Breusch-Godfrey (BG) Test

To avoid the pitfalls of the Durbin Watson *d*-test, Breusch and Godfrey have proposed a test criterion for autocorrelation that is general in nature. This is in the sense that:

(a) It can handle non-stochastic regressors as well as the lagged values of Y_t ;

(b) It can deal with higher-order autoregressive schemes such as AR(2), AR(3) … etc.

(c) It can also handle simple or higher order moving averages.

The BG-Test is also referred to as the LM (Lagrange Multiplier) Test (see Unit 8). Let us now consider a two-variable regression model to see how the BG test works.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad \ldots (12.7)$$

where u_t follows a P^{th} order auto regressive scheme AR(P) like:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \ldots + \rho_\rho u_{t-p} + v_t \quad \ldots (12.8)$$

where v_t is the white noise or the stochastic error term. We wish to test:

$$H_0: \rho_1 = \rho_2 = \ldots \rho_p = 0 \quad \ldots (12.9)$$

The null hypothesis says that there is no autocorrelation of any order. Now, the BG test involves the following steps:

(a) Estimate the model $Y_t = \beta_1 + \beta_2 X_t + u_t$ by OLS method and obtain the residuals e_t.

(b) Regress the residuals e_t on the *p*-lagged values of estimated residuals obtained in step (a) above, i.e., $e_{(t-1)}$, $e_{(t-2)}$, ..., $e_{(t-p)}$ [as in equation (12.8)]. Here we take the residual e_t which are estimate of the error u_t, as the error term is not known.

(c) Obtain R^2 from the auxiliary regression (12.8) in the step (b) above.

(d) Now, for large samples, the Breusch and Godfrey test statistic is computed as:

$$(n-p)R^2 \sim \chi_p^2 \quad \ldots (12.10)$$

It is called LM test, as it has a similar form to the LM test described in Unit 8. The BG test statistic follows chi-squares distribution with p degrees of freedom where *p* is the number of regressors in the auxiliary regression (equation (12.8)).

We draw inferences from the BG test as follows:

(i) If $(n-p)R^2 > \chi^2_{critical}$, we reject H_0 and conclude that at least one ρ is statistically different from zero, i.e., there exists autocorrelation.

(ii) If $(n-p)R^2 < \chi^2_{critical}$, we do not reject H_0 and conclude that there exists no autocorrelation.

Check Your Progress 2 [Answer the questions in 50-100 words within the space given]

1) State the methods of detecting autocorrelation.

……

……

……

……

……

2) Specify the test statistic applied in the DW test.

……

……

……

……

……

3) State the assumptions under which the DW test is valid.

……

……

……

……

……

4) Point out the limitations of the DW test.

……

……

……

……

……

5) In what ways the BG test for autocorrelation is an improvement over the DW test?

……

……

……

……

…….

12.6 REMEDIAL MEASURES FOR AUTOCORRELATION

To suggest remedial measures for autocorrelation, we assume the nature of inter-dependence in the error term u_t in a regression model like:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \quad \ldots (12.11)$$

and that the error term is following an AR (1) scheme like:

$$u_t = \rho u_{t-1} + v_t \qquad -1 \leq \rho \leq 1 \quad \ldots (12.12)$$

where v_t is assumed to follow the OLS assumptions. We first consider the case where ρ is known. Here, transforming the model in a certain manner (called as the Cochrane Orcutt procedure) will reduce the equation to an OLS compatible model. When ρ is not known, we need some simple approaches which help us in overcoming the situation of autocorrelation. Let us study these approaches now.

12.6.1 Autoregressive Scheme is Known: Cochrane-Orcutt Transformation

Suppose we know the value of ρ. This helps us to transform the regression model given at (12.11) in a manner that the error term becomes free from autocorrelation. Subsequently, we apply the OLS method to the transformed model. For this, we consider a one-period lag in (12.11) as:

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \quad \ldots (12.13)$$

Let us multiply equation (12.13) on both the sides by ρ. We obtain:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \quad \ldots (12.14)$$

Let us now subtract equation (12.14) from equation (12.11) to obtain:

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + v_t \quad \ldots (12.15)$$

Note that we have used v_t for the new disturbance term above. Let us now denote:

$$Y_t^* = (Y_t - \rho Y_{t-1})$$

$$X_t^* = (X_t - \rho X_{t-1})$$

$$\beta_1^* = \beta_1(1-\rho)$$

The transformed model will be

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + v_t \quad \ldots (12.16)$$

Now, the transformed variables $Yt^*$and Xt^*will have the desirable BLUE property. The estimators obtained by applying the OLS method to (12.16) are called the Generalized Least Squares (GLS) estimators. The transformation as suggested above is known as the Cochrane-Orcutt transformation procedure.

12.6.2 Autoregressive Scheme is not Known

Suppose we do not know ρ. Thus, we need methods for estimating ρ. We first consider the case where ρ = 1. This amounts to assuming that the error terms are perfectly positively autocorrected. This case is called as the First Difference Method. If this assumption holds, a generalized difference equation can be considered by taking the difference between (12.11) and its first order autoregressive schemes as:

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + v_t \quad \ldots (12.17)$$

i.e., $\Delta Y_t = \beta_2 \Delta X_t + v_t \quad \ldots (12.18)$

where the symbol Δ (read as delta) is the first difference operator. Note that the difference model (12.17) has no intercept. If ρ is not known, then we can estimate ρ by the following two methods.

(i) Durbin Watson Method

From equation (12.6) we see that *d*-statistic and ρ are related. We can this relationship to estimate ρ. The *d*-statistic and ρ are related as:

$$\rho \approx 1 - \frac{d}{2} \quad \ldots (12.19)$$

If the value of *d* is known, then $\hat{\rho}$ can be estimated from the *d*-statistic.

((ii) The OLS Residuals (e_t) Method

Here, we consider the first order autoregression scheme as in (12.12), i.e.,

$u_t = \rho u_{t-1} + v_t$. Since u_t is not directly observable, we use its sample counterpart e_t and run the following regression:

$$e_t = \hat{\rho} e_{t-1} + v_t \quad \ldots (12.20)$$

Note that $\hat{\rho}$ is an estimator of ρ. In small samples, $\hat{\rho}$ is a biased estimator of ρ. As sample size increases, the bias disappears.

12.6.3 Iterative Procedure

This is also called as the Cochrane-Orcutt iterative procedure. We consider the two variable model with the AR(1) scheme for autocorrelation as discussed earlier. That is, we consider: $Y_t = \beta_1 + \beta_2 X_t + u_t$ where $u_t = \rho u_{t-1} + v_t$ with $-1 \leq \rho \leq 1$. We have taken only one explanatory variable for simplicity but we can have more than one explanatory variable too. The iterative procedure suggested by Cochrane-Orcutt has the following steps:

(i) Estimate the equation $u_t = \rho u_{t-1} + v_t$ by the usual OLS method.

(ii) From the above, obtain the residuals e_t.

(iii) Using the residuals e_t, run the regression $e_t = \hat{\rho}\, e_{t-1} + v_t$ and obtain $\hat{\rho}$.

(iv) Use $\hat{\rho}$ obtained in (iii) above to multiply the equation $u_t = \rho u_{t-1} + v_t$.

(v) Now, obtain the generalized difference equation as:

$Y_t^* = \beta_1^* + \beta_2 X_t^* + e_t$ where, $Y_t^* = Y_t - Y_{t-1}$, $X_t^* = X_t - \rho X_{t-1}$ and

$\beta_1^* = \beta_1(1 - \hat{\rho})$

(vi) We are not sure that $\hat{\rho}$ estimated in (iii) above is the best estimate of ρ. Therefore, we repeat the steps (ii) and (iii) to obtain the new residuals e_t^*.

(vii) Now estimate the regression $e_t^* = \hat{\rho} e_{t-1}^* + w_t$to obtain the new estimate of $\hat{\rho}$.

We thus obtain the second-round estimate of ρ. Since we are not sure if the second round estimate of ρ is the best, we go for the third round estimate and so on. We repeat the same steps again and again. Due to this repetitive steps followed, this procedure, suggested by Cochrane-Orcutt, is called the 'iterative procedure'. We stop the iteration when the successive estimates of ρ differ by a small amount (less than 0.01 or 0.005).

12.7 LAGGED DEPENDENT VARIABLE

The Durbin-Watson method is not applicable when the regression model includes lagged value of the dependent variable as one of the explanatory variables. In such models, the *h*-statistic suggested by Durbin is used to identify the presence of autocorrelation in the regression model. Let us consider the regression model as:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + v_t \quad \ldots (12.21)$$

In equation (12.21), we have two explanatory variables: X_t and Y_{t-1} with Y_{t-1} as a lagged dependent variable (with one-period lag). For equation (12.21) the *d*-statistic is not applicable to detect autocorrelation. For such models, Durbin suggests replacing the *d*-statistic by the *h*-statistic taken as:

$$h \approx \hat{\rho} = \sqrt{\frac{n}{1 - n\, Var(b_3)}} \quad \ldots (12.22)$$

where, *n* = sample size, $\hat{\rho}$ = the estimator of the autocorrelation coefficient, and

$var(b_3)$ = variance of estimator of β_3, the lagged dependent variable in (12.21).

The null hypothesis is H_0: $\rho = 0$. Durbin has shown that for large samples the *h*-statistic is distributed as h~$N(0,1)$. For normal distribution, we know that the critical value at 5 per cent level of significance is 1.96 and at 1 per cent level of significance it is 2.58. Using this information, we can draw inference from equation (12.22) as follows:

(i) If the computed value of h is greater than the critical value of h, we reject H_0. We interpret the result as existence of no autocorrelation.

(ii) If the computed value of h is less than the critical value of h, we do not reject H_0. We interpret the result as existence of autocorrelation.

Check Your Progress 3 [Answer the questions in 50-100 words within the space given]

1) Outline the transformation procedure suggested by Cochrane-Orcutt to resolve the problem of autocorrelation.

……

……

……

……

……

2) State how the iterative procedure of Cochrane-Orcutt is applied in the case of autocorrelation in a dataset. Why is it called iterative procedure?

……

……

……

……

……

3) What is the advantage of using the h-statistic in regression model having autocorrelation problem?

……

……

……

……

……

11.8 LET US SUM UP

The unit has discussed the concept of autocorrelation in regression models. The consequences of the presence of autocorrelation, its detection and techniques that provide remedial measures for such situations have been explained. The unit also discusses the case of autocorrelation in regression models with lagged dependent variables.

11.9 ANSWERS/ HINTS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) Autocorrelation refers to the presence of correlation between the error terms of any two observations. This means if U_i and U_j are the error terms, then *Corr (Ui, Uj)* $\neq 0$ for $i \neq j$. In the CLRM, one of our assumptions is that the *Corr (Ui, Uj)* $= 0$. This means the two error terms are not correlated. Violation of this assumptions is a situation of autocorrelation.

2) The problem of autocorrelation is more common in time series data. This is because a phenomena affecting the error term in one point of time is more likely to be influencing the error term in the next point of time. This is especially identified as the factor of 'inertia or sluggishness'. Across units of cross section this is less likely. But it cannot be ruled out even in cross section data. In such cases, due to the spatial effect in cross section data, which is more like a demonstration effect, it is distinctly termed as spatial correlation.

3) Inertia or sluggishness, specification error in the model, cobweb phenomenon and data smoothening.

4) The consequences are: (i) least squares estimators are not efficient, (ii) the estimated variances of OLS estimates are biased, (iii) the standard error of true variances are underestimated, (iv) we are more likely to commit an error in deciding on the hypothesis of 'no statistical significance' of a particular estimated coefficient i.e. the decisions based on t and F tests would be unreliable, (v) estimated error variance would be biased and (vi) the value of R^2 would be misleading or unreliable.

Check Your Progress 2

1) Time sequence plotting (graphical method), Durbin-Watson test and Breusch-Godfrey (BG) Test.

2) $\boldsymbol{d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}}$. It is the ratio of the sum of the squared differences in the successive residuals to the residual sum of squares.

3) The regression model includes an intercept term, the X variables are non-stochastic, the error term follows the following mechanism $u_t = \rho u_{t-1} + v_t$, $-1 \leq \rho \leq 1$, and the regression does not contain any lagged values of the dependent variable as one of the explanatory variables.

4) The one drawback of the d-test is that it has two zones of indecision, viz., $d_L < d < d_U$ and $(4 - d_U < d < 4 - d_L)$.

5) (i) It can handle non-stochastic regressors as well as the lagged values of Y_t, (ii) it can deal with higher-order autoregressive schemes such as AR(2)… etc. and (iii) it can also handle simple or higher order moving averages.

Check Your Progress 3

1) In this method we lag the regression equation by one period; multiply it by $\boldsymbol{\rho}$; and subtract it from the original regression equation. This gives us a transformed regression model. When estimated by OLS method, the estimators of the transformed model are BLUE.

2) In Sub-Section 12.6.3 we have outlined steps of the Cochrane-Orcutt iterative procedure. You should go through it and answer.

3) The *h*-statistic can be used in regression models having lagged dependent variables as explanatory variables.

UNIT 13 MODEL SELECTION CRITERIA*

Structure

13.0 OBJECTIVES

After going through this unit, you will be able to

- appreciate the importance of correct specification of an econometric model;
- identify the important issues in specification of econometric models;
- find out the consequences of including an irrelevant variable;
- find out the consequences of excluding a relevant variable; and
- find out the impact of measurement errors in dependent and independent variables.

13.1 INTRODUCTION

In the previous Units of the course we have discussed about various econometric tools. We began with the classical two variable regression model. Later on, we extended it to the classical multiple regression model. The steps of carrying out the ordinary least squares (OLS) method were discussed in details. Recall that the

*Dr. Sahba Fatima, Independent Researcher, Lucknow.

classical regression model is based on certain assumptions. When these assumptions are met, the OLS estimators are the best linear unbiased estimators (BLUE). When these assumptions are violated the OLS estimators are not BLUE – they lose some of their desirable properties. Therefore, when some of the classical assumptions are not fulfilled, we have to adopt some other estimation method.

Thus far our objective has been to explain how various estimation methods are applied. Now let us look into certain other important issues regarding specification of econometric models.

13.2 ISSUES IN SPECIFICATION OF ECONOMETRIC MODEL

A model refers to a simplified version of reality. It allows us to explain, analyse and predict economic behavior. An economic model can be for a microeconomic agent such as household or firm. In macroeconomics, it represents the behavior of the economy as a whole. In economic models we identify relevant economic variables (such as income, output, expenditure, investment, saving, exports, etc.) and establish relationship among them. The relationships among these variables may be expressed through diagrams or mathematical equations. There could be economic models without mathematical expressions, but such models may not be precise.

Recall from Unit 1 of this course that there are eight steps to be followed in an econometric study. The first three steps are as follows:

(i) Construction of a statement of theory or hypothesis

(ii) Specification of mathematical model of the theory

(iii) Specification of econometric model

Based on economic theory or logic we construct the hypothesis. We specify the hypothesis in mathematical terms. Further, we add a stochastic error term (u_i) to transform it into an econometric model. We decide on the estimation method (such as OLS, GLS, maximum likelihood, etc.) subsequently.

13.2.1 Model Specification

While building an econometric model we first consider the logic or theory behind the model. The empirical or methodological considerations come later. The accuracy of the estimated parameters and the inferences drawn from the model depend upon the correct specification of the model.

An econometric model comprises a dependant variable, independent variable(s) and the error term. The dependant variable should be logically explained by the independent variables. Next is the functional form of the regression model, which should be specified correctly.

Let me illustrate the point through an example. In the case of a firm, we assume that there are two factors of production, viz., capital and labour. We club all types of labour into a homogeneous category – we do not distinguish between a manager and a worker in the field! Thus you should remember that we ignore the details and concentrate on the major issues in a model. Secondly, we assume that the production function takes a particular form, say Cobb-Douglas. But, remember that it is just an assumption! The production function in reality could be of some other form. Thus we have to logically explain the functional form (regression equation) of the model.

Regression analysis derives its robustness from the assumption that the econometric model under study is correctly specified. In Unit 4 of this course we specified the assumptions such that the econometric model must bring efficient estimates of the parameters in the model. Ordinary Least Squares (OLS) method is based on the assumption that regression model is correctly specified. Correct specification has three important elements:

a) all the necessary independent variables are included in the model,

b) no redundant variable IS included in the model, and

c) the model is specified using the correct functional form.

13.2.2 Violation of Basic Assumptions

An economic model is based on certain assumptions. Recall that we made the following assumptions regarding the multiple regression model (see Unit 7):

a) The regression model is linear in parameters

b) $E(X_i u_i) = 0$ (regressor is non-stochastic)

c) $E(u_i) = 0$

d) $E(u_i)^2 = \sigma^2$

e) $E(u_i u_j) = 0$ for $i \neq j$

f) The explanatory variables (X_i) are independent of one another.

Let us look into the implications of the above assumptions. Assumption (a) says that the regression model is linear in parameters. Standard regression model usually takes the following form

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad \ldots (13.1)$$

Equation (13.1) is linear in parameters (there are no such terms as β_i^2, for example) and linear in variables. Examples of non-linear regression models are logarithmic functions, logistic functions, trigonometric functions, exponential functions, etc. For estimation of non-linear models, the OLS method cannot be applied.

Assumption (b) says that X_i and u_i are independent. Thus if we take the X_i values randomly, the joint probability of both that X_i and u_i will not be zero. In order to avoid this problem we assume that X_i is non-stochastic. All explanatory variables are fixed in repeated sampling.

Assumption (c) says that the mean of the error term (u_i) is zero. There could be errors in individual observations; on the whole these errors cancel out. If $E(u_i) \neq 0$, OLS estimator of the intercept term (β_1) will be biased. Estimators of the slope parameters β_2 and β_3 will remain unbiased. For example, suppose $E(u_i) = 3$. In that case $E(Y_i)$ will be

$$E(Y_i) = E(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i)$$

Remember that β_i are parameters of the model. They are constants. We have assumed X_i to be fixed across samples. Thus

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + E(u_i) \quad \ldots (13.2)$$

If $E(u_i) = 3$, we can say that

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + 3$$

Thus the intercept term will be $(\beta_1 + 3)$. Remember that if assumption (d) is violated we have the problem of heteroscedasticity, which is discussed in Unit 11. If assumption (e) is violated we have the problem of autocorrelation, that we have discussed in Unit 12. In case the assumption (f) is violated we have the problem of multicollinearity (see Unit 10).

Check Your Progress 1

1) List the assumptions of the classical regression model.

.....

.....

.....

.....

.....

2) Do you agree that correct specification of an econometric model is important? Why?

……

……

……

……

……

3) What are the implications of violations of the basic assumptions classical regression model?

.....

.....

.....

.....

4) List three types of specification error that we encounter in an econometric model.

.....

.....

.....

.....

.....

13.3 CONSEQUENCES OF SPECIFICATION ERRORS

As pointed out earlier, we usually encounter three kinds of problems in an econometric model:

a) Inclusion of irrelevant/redundant variables

b) Omission of relevant variables

c) Incorrect functional form of the model

Each of the above problem results in a different kind of bias. We discuss each of these problems below.

13.3.1 Inclusion of Irrelevant Variable

Let us consider the case where some irrelevant variable is included in the regression model. Suppose the true model is

$Y_i = \beta_0 + \beta_1 X_{1i} + u_i$ … (13.3)

But we somehow include a redundant variable, i.e., we estimate the following equation:

$Y_i = \beta_{0s} + \beta_{1s} X_{1i} + \beta_{2s} X_{2i} + v_i$ … (13.4)

For the true model (13.3), the slope coefficient is expressed as

$$\hat{\beta}_1 = \frac{\sum y x_1}{\sum x_1^2} \quad \text{… (13.5)}$$

which is unbiased.

For the model (13.4) that we have taken, we obtain

$$\tilde{\beta}_1 = \hat{\beta}_{1s} = \frac{(\sum yx_1)(\sum x_2^2)-(\sum yx_2)(\sum x_1x_2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.6)$$

Now the true model in deviation form is

$$y_i = \beta_1 x_1 + (u_i - \bar{u}) \quad \ldots (13.7)$$

Substituting for y_i from (13.7) into (13.6) and simplifying, we obtain

$$E(\tilde{\beta}_1) = E(\hat{\beta}_{1s}) = \beta_1 \frac{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.8)$$

From equation (13.8) we find that

$E(\tilde{\beta}_1) = \beta_1$

Thus, inclusion of an irrelevant variable provides us with unbiased estimator of β_1.The estimator of the redundant variable $\hat{\beta}_{2s}$ is given by

$$\hat{\beta}_{2s} = \frac{(\sum yx_2)(\sum x_1^2)-(\sum yx_1)(\sum x_1x_2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.9)$$

If we substitute for y_i from (13.7) in (13.9) and re-arrange terms, we obtain

$$E(\tilde{\beta}_2) = E(\hat{\beta}_{2s}) = \beta_2 \frac{(\sum x_1x_2)(\sum x_1^2)-(\sum x_1x_2)(\sum x_1^2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1x_2)^2} \quad \ldots (13.10)$$

Thus, $E(\tilde{\beta}_2) = E(\hat{\beta}_{2s}) = 0$

So, we find that $\hat{\beta}_{2s}$ which is absent from the true model has its coefficient 0. Thus we obtain unbiased estimators for both the parameters.

This leads us to conclude that inclusion of irrelevant variables is not that harmful as omission of relevant variables. As an extra variable is added to the model, we observe that there is an increase in R-squared. The variance of the parameters will not be efficient.

Therefore, the specification error in the nature of inclusion of irrelevant variables in the model, will produce unbiased but inefficient least squares estimators of the parameters. The larger variance reduces the precision of the estimates resulting in wider confidence intervals. This may lead to type II error (the error of not rejecting a null hypothesis when the alternative hypothesis is actually true).

13.3.2 Omission of Relevant Variable

Now let us look into the other side of the spectrum – excluding a relevant variable. Since a relevant variable is not included in the model (although it influences the dependent variable) its impact will be included in the residuals. As a result, the residuals will show a systematic pattern rather than being white noise as required by Gauss-Markov theorem. Also, the coefficient of the included variable will be biased.

Suppose the true equation (in deviation form) is

$$y = \beta_1 x_1 + \beta_2 x_2 + u \quad \ldots (13.11)$$

Instead of estimating equation (13.11) suppose we omitted x_2. The following equation is estimated,

$y = \beta_1^* x_1 + e$ … (13.12)

Equation (13.12) is a case of omitted variable, and hence incorrect model specification. In the model with omitted variable (incorrect model) the estimate of β_1^* is

$$\hat{\beta}_1^* = \frac{\sum x_1 y}{\sum x_1^2} \quad \text{… (13.13)}$$

In order to calculate the bias in the estimated value of β_1 in the incorrect model (equation (13.12)) as compared to the true model (equation (13.11)), we take the following steps:

Substituting the expression of y from the true model in (13.11), we get

$$\hat{\beta}_1^* = \frac{\sum x_1(\beta_1 x_1 + \beta_2 x_2 + u)}{\sum x_1^2} = \beta_1 + \beta_2 \frac{\sum x_1 x_2}{\sum x_1^2} + \frac{\sum x_1 u}{\sum x_1^2} \quad \text{… (13.14)}$$

Since E $(\sum x_1 u) = 0$ we get

$$E(\hat{\beta}_1^*) = \beta_1 + b_{21}\beta_2 \quad \text{… (13.15)}$$

where $b_{21} = \frac{\sum x_1 x_2}{\sum x_1^2}$ is the regression coefficient from a regression of X_2 (omitted variable) on X_1.

Thus $\hat{\beta}_1^*$ is a biased estimator for β_1 and the bias is given by

Bias = (coefficient of the excluded variable) × (regression coefficient in a regression of the excluded variable on the included variable) … (13.16)

In the deviation form, the three-variable population regression model can be written as

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + (u_i - \bar{u}) \quad \text{… (13.17)}$$

First multiplying by x_2 and then by x_3, the usual normal equations are

$$\sum y_i x_{2i} = \beta_2 \sum x_{2i}^2 + \beta_3 \sum x_{2i} x_{3i} + \sum x_{2i} (u_i - \bar{u}) \quad \text{… (13.18)}$$

$$\sum y_i x_{3i} = \beta_2 \sum x_{2i} x_{3i} + \beta_3 \sum x_{3i}^2 + \sum x_{3i} (u_i - \bar{u}) \quad \text{… (13.19)}$$

Dividing (13.18) by $\sum x_{2i}^2$ on both sides, we obtain

$$\frac{\sum y_i x_{2i}}{\sum x_{2i}^2} = \beta_2 + \beta_3 \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \quad \text{… (13.20)}$$

Thus we have

$$b_{y2} = \frac{\sum y_i x_{2i}}{\sum x_{2i}^2}$$

$$b_{32} = \frac{\sum x_{2i} x_{3i}}{\sum x_{2i}^2}$$

Hence (13.20) can be written as

$$b_{y2} = \beta_2 + \beta_3 b_{32} + \frac{\sum x_{2i}(u_i - \bar{u})}{\sum x_{2i}^2} \quad \ldots (13.21)$$

Taking the expected value of (13.21) we obtain

$$E(b_{y2}) = \beta_2 + \beta_3 b_{32} \quad \ldots (13.22)$$

Similarly, if x_2 is omitted from the model, the bias in $E(b_{y3})$ can be calculated.

The variance of β_1^* (parameter of the incorrect model) can also be derived by using the formula for variance. As it is a bit complex, we do not present it here. You should note that the variance of β_1^* is higher than that of β_1. An implication of the above is that usual tests of significance concerning parameters are invalid, if some of the relevant variables are excluded from a model.

Thus we know that

(i) When an irrelevant variable is included in the model: (a) the estimators of parameters are unbiased, (b) efficiency of the estimators decline, and (c) estimator of the error variance is unbiased. Thus conventional tests of hypothesis are valid. The inferences drawn could be somewhat erroneous.

(ii) When a relevant variable is dropped from the model: (a) estimators of parameters are biased, (b) efficiency of estimators decline, and (c) estimator of error variance is biased. Thus conventional tests of hypothesis are invalid. The inferences drawn are faulty.

13.3.3 Incorrect Functional Form

Apart from inclusion of only relevant variables in an econometric model, another specification error pertains to functional form. There is a tendency the part of researchers to assume a linear relationship between variables. This however is not always true. If the true relationship is non-linear and we take a linear regression model for estimation, we will not be able to draw correct inferences. There are test statistics available to choose among functional forms. We will discuss these test statistics in Unit 14.

Check Your Progress 2

1) Explain the consequences of inclusion of an irrelevant variable.

……

……

……

……

……

2) Explain the consequences of excluding a relevant variable.

.....

.....

.....

.....

.....

.....

13.4 ERROR OF MEASUREMENT IN VARIABLES

So far we have assumed the variables in the econometric model under study are measured correctly. It means that there are no measurement errors in both explained and explanatory variables. Sometimes we do not have data on the variables that we want to use in the model. This could be for various reasons such as non-response error, reporting error, and computing error. A classic example of measurement error pertains to the variable permanent income used in the Milton Friedman model. Measurement error in variables is a serious problem in econometric studies. There are two types of measurement errors:

(i) Measurement error in dependent variable, and

(ii) Measurement error in independent variable.

13.4.1 Measurement Error in Dependent Variable

Let us consider the following model:

$Y_i^* = \alpha + \beta X_i + u_i$ … (13.23)

where Y_i^* is permanent consumption expenditure

X_i is current income, and

u_i is the stochastic disturbance term.

(we place a star mark (*) on the variable that is measured with errors)

Since Y_i^* is not directly measureable, we may use an observable expenditure variable Y_i such that

$Y_i = Y_i^* + e_i$ … (13.24)

where e_i denote measurement error in Y_i^*.

Therefore, instead of estimating

$Y_i^* = \alpha + \beta X_i + u_i$, we estimate

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

$= \alpha + \beta X_i + (u_i + e_i)$

Let us re-write the above equation as

$$Y_i = \alpha + \beta X_i + v_i \quad \ldots (13.25)$$

where $v_i = u_i + e_i$

In equation (13.25) we take v_i as a composite error term comprising population disturbance term (u_i) and measurement error term (e_i).

Let us assume that the following classical assumptions hold

a) $E(u_i) = E(e_i) = 0$

b) Cov $(X_i, u_i) = 0$

c) Cov $(u_i, e_i) = 0$

An implication of (c) above is that the stochastic error term and the measurement error term are uncorrelated. Thus expected value of the composite error term is zero; $E(v) = 0$. By extending the logic given in Unit 4, we can say that $E(\hat{\beta}) = \beta$. It implies that $\hat{\beta}$ is *unbiased*.

Now let us look into the issue of variance in the case of measurement error in the dependent variable. As you know, variance of the estimator $\hat{\beta}$ in a two variable regression model (13.23) is given by

$$\text{Var}(\hat{\beta}) = \frac{\sigma_u^2}{\sum x_i^2},$$

For the composite error term, this will translate into

$$\text{Var}(\hat{\beta}) = \frac{\sigma_u^2 + \sigma_e^2}{\sum x_i^2} = \frac{\sigma_v^2}{\sum x_i^2} \quad \ldots (13.26)$$

Thus we see that the variance of the error term is larger if there is measurement error in the dependent variable. This leads to inefficiency of the estimators. They are not best linear unbiased estimators (BLUE).

13.4.2 Measurement Error in Independent Variable

There could be measurement error in explanatory variables. Let us assume the true regression model to be estimated is

$$Y_i = \alpha + \beta X_i^* + u_i \quad \ldots (13.27)$$

Suppose we do not have data on variable X_i^*. On the other hand, suppose we have data on X_i. In that case, instead of observing X_i^*, we observe

$$X_i = X_i^* + w_i \quad \ldots (13.28)$$

where w_i represents error of measurement in X_i^*.

In the permanent income hypothesis model, for example,

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

where Y_i is current consumption expenditure

X_i^* is permanent income

u_i is stochastic disturbance term (equation error)

From equation (13.27) and (13.28) we find that

$$Y_i = \alpha + \beta(X_i - w_i) + u_i \quad \ldots (13.29)$$

$$= \alpha + \beta X_i + (u_i - \beta w_i)$$

$$= \alpha + \beta X_i + z_i \quad \ldots (13.30)$$

where $z_i = (u_i - \beta w_i)$. You should notice that z_i is made up of two terms: stochastic error and measurement error.

Now, let us assume that w_i has zero mean; it is serially independent; and it is uncorrelated with u_i. Even in that case, the composite error term z_i is not independent of the explanatory variable X_i.

$$\text{Cov}(z_i, X_i) = \text{E}[z_i - E(z_i)[X_i - E(X_i)]$$

$$= \text{E}(u_i - \beta w_i)(w_i)$$

$$= \text{E}(-\beta w_i^2)$$

$$= -\beta \sigma_w^2 \quad \ldots (13.31)$$

From (13.31) we find that the independent variable and the error term are correlated. This violates the basic assumption of the classical regression model that the explanatory variable is uncorrelated with the stochastic disturbance term. In such a situation the OLS estimators are not only biased but also inconsistent, that is they remain biased even if the sample size *n* increases infinitely.

Check Your Progress 3

1) Explain the consequences measurement error in the dependent variable.

.....

.....

.....

.....

.....

.....

2) Explain the consequences of measurement error in the explanatory variable.

.....

.....

.....

.....

.....

.....

3) Measurement error in the dependent variable is a lesser evil than measurement error in the explanatory variable.

.....

.....

.....

.....

.....

.....

13.5 LET US SUM UP

Correct specification of an econometric model determines the accuracy of the estimates obtained. Therefore, correct specification of an econometric model is very important. Economic theory and logic guide us in specification of econometric models.

In order to correctly specify an econometric model all relevant explanatory variables should be included in the model. No relevant explanatory variable should be excluded from the model. Further, the functional form of the model should be correct.

At times we do not get appropriate variable required in an econometric model. In such cases there could be cases where either dependent variable or independent variable is measured with certain error. Measurement error in dependent variable is a lesser evil than the measurement error in the independent variable.

13.6 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) The basic assumptions of the classical regression model are as follows:

   a) The regression model is linear in parameters

   b) $E(X_i u_i) = 0$ (regressor is non-stochastic)

   c) $E(u_i) = 0$

   d) $E(u_i)^2 = \sigma^2$

   e) $E(u_i u_j) = 0$ for $i \neq j$

   f) The explanatory variables (X_i) are independent of one another.

2) Go through Section 13.2. It is important because incorrect specification has serious implications on desirable properties of the estimators.

3) Go through Sub-Section 13.2.2 and answer.

4) The important specific issues are: inclusion of irrelevant/redundant variables; omission of relevant variables; and incorrect functional form of the model

Check Your Progress 2

1) The estimator is unbiased but inefficient. See Sub-Section 13.3.1.

2) The estimator is biased as well as inefficient. See Sub-Section 13.3.2.

Check Your Progress 3

1) Go through Sub-Section 13.4.1 and answer.

2) Go through Sub-Section 13.4.2 and answer.

3) If there is measurement error in dependent variable the estimator is unbiased but inefficient. Measurement error in explanatory variable results in biased estimator. See Section 13.4 for details.

UNIT 14 TESTS FOR SPECIFICATION ERROR*

Structure

14.1 INTRODUCTION

In the previous Unit we highlighted the consequences of specification errors. There could be three types of specification errors; inclusion of an irrelevant variable, exclusion of a relevant variable, and incorrect functional form. When the econometric model is not specified correctly, the coefficient estimates, the confidence intervals, and the hypothesis tests are misleading and inconsistent. In view of this, econometric models should be correctly specified.

While building a model we face a lot of difficulties in specifying a model correctly. In some cases economic theory is quite transparent about the dependent variables and the independent variables. In some other cases still it is in a hypothesis stage. Researchers are still working in that area to confirm the hypothesis suggested by others. In such cases, what we have a dependent variable and a set of explanatory variables. Out of these explanatory variables we have to select the most appropriate ones.

* Dr. Sahba Fatima, Independent Researcher, Lucknow.

Econometric theory suggests certain criteria and test statistics. On the basis of these criteria we select the most appropriate econometric model. We describe some of these criteria below.

14.2 OBJECTIVES

After going through this Unit, you should be in a position to

- identify econometric models that are not specified correctly;
- take remedial measures for correcting the specification error; and
- evaluate the performance of competing models.

14.3 TESTS FOR IDENTIFYING THE MOST EFFICIENT MODEL

As pointed out above, econometric models should be specified correctly. Any spurious relationship should be identified and excluded from the model. There are certain tests for this purpose. These tests can be used under specific circumstances in conjunction with practical understanding of the variables and an enlightened study of it through the related literature. Following tests are most commonly used for model testing and evaluation.

14.3.1 The R^2 Test and Adjusted-R^2 Test

We have discussed the concept of coefficient of determination (R^2) in Unit 4. As you know, the coefficient of determination indicates the explanatory power of a model. If, for example, $R^2 = 0.76$ we can infer that 76 per cent variation in the dependent variable is explained by the explanatory variable in the model.

We define R^2 as follows:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} \quad \text{... (14.1)}$$

where TSS = Total Sum of Squares

ESS = Explained Sum of squares

RSS = Residual Sum of Squares

As you know,

$$TSS = RSS + ESS \quad \text{... (14.2)}$$

Dividing both sides of equation (14.2) by TSS, we find that

$$\frac{RSS}{TSS} + \frac{ESS}{TSS} = 1 \quad \text{... (14.3)}$$

Since $R^2 = \frac{ESS}{TSS}$, we observe that R^2 lies between 0 and 1 necessarily. Its closeness to 1 indicates better fit of the model. If R^2 is close to one, RSS is much smaller compared to ESS. Therefore, very little residual will be left. Thus a

model with higher R^2 is preferred. You should however keep in mind that a very high R^2 indicates the presence of multicollinearity in the model. If the R^2 is high but the t-ratio of the coefficients are not statistically significant you should check for multicollinearity. The R^2 is calculated on the basis of the sample data.

Thus the explanatory variables included the model are considered for estimation of R^2. Variables not included in the model do not account for the variation in the dependent variable.

There is a tendency of the R^2 to increase if more explanatory variables are added. Thus, we are tempted to add more explanatory variables to increase the explanatory power of the model. If we add irrelevant explanatory variables in a model, the estimators are unbiased, but there is an increase in the variance of the estimators. This makes forecast and analysis on the basis of such models unreliable.

In order to overcome this difficulty, we use the 'adjusted-R^2'. It is denoted by $\bar{R}^2$ and defined as follows:

$$\bar{R}^2 = 1 - \frac{ESS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} \quad \ldots (14.4)$$

where n is the number of observations and k is the number of regressors. As you know the TSS has a degree of freedom of $(n-1)$ while the ESS has a degree of freedom of $(n-k)$. Thus, $\bar{R}^2$ takes into account the degrees of freedom of the model. The $\bar{R}^2$ penalises the addition of explanatory variables. It is observed that there is an increase in $\bar{R}^2$ only if the t-value (absolute number) of the additional explanatory variable is greater than 1. Hence, superfluous variables can be identified and eliminated from the model. The restriction here is to regress all the independent variable against the same dependent variable.

Remember that we can compare the $\bar{R}^2$ of two models only if the dependent variable is the same. For example, we cannot compare two models if in one model the explanatory variable is Y and in the other model the explanatory variable in *log*Y.

14.3.2 Akaike Information Criterion (AIC)

Another method for identifying the mis-specification in a model is Akaike Information Criterion (AIC). This method also penalises the addition of regressors as we can see from the formula below:

$$AIC = e^{2k/n}\sum\frac{\hat{u}_i^2}{n} = e^{2k/n}\frac{RSS}{n} \quad \ldots (14.5)$$

where k is the number of regressors (explanatory variables) and n is the number of observations.

We can further simplify equation (14.5) as

$$\ln AIC = \left(\frac{2k}{n}\right) + \ln\left(\frac{RSS}{n}\right) \quad \ldots (14.6)$$

where ln AIC is the natural log of AIC, and $\frac{2k}{n}$ is the penalty factor.

Remember that the model with a lower value of lnAIC is considered to be better. Thus, when we compare two models by using the AIC criterion, the model with lower value of AIC has a better specification. The logic is simple. An econometric model that reduces the residual sum of squares is a better specified model.

14.3.3 Schwarz Information Criterion

The Schwarz Information Criterion (SIC) also relies on the RSS, like the AIC criterion mentioned above. This method also is popular for analysing correct specification of an econometric model. The SIC is defined as follows:

$$SIC = n^{k/n} \frac{\sum \hat{u}^2}{n} = n^{k/n} \frac{RSS}{n} \quad \ldots (14.7)$$

If we take in log-form, equation (14.7) is given as

$$\ln SIC = \frac{k}{n} \ln n + \ln\left(\frac{RSS}{n}\right) \quad \ldots (14.8)$$

where $[(k/n) \ln n]$ is the penalty factor. Note that the SIC criterion imposes a harsher penalty for inclusion of explanatory variable compared to the AIC criterion.

14.3.4 Mallow's C_p Criterion

When we do not include all the relevant variables in a model, the estimators are biased. The Mallow's C_p Criterion evaluates such bias to find out whether there is significant deviation from the unbiased estimators. Thus, the Mallow's C_p Criterion helps us in selecting the best among competing econometric models.

If some of the explanatory variables are dropped from a model, there is an increase in the residual sum of squares (RSS). Let us assume that the true model has k regressors. For this model, $\hat{\sigma}^2$ is the estimator of true σ^2. Now, suppose we drop p regressors from the model. The residual sum of squares obtained from the truncated model is RSS_p. The Mallow's C_p Criterion is based on the following formula:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \quad \ldots (14.9)$$

where n is the number of observations.

While choosing a model according to the C_p criterion, the model with the lowest C_p value is preferred.

14.4 CAUTION ABOUT MODEL SELECTION CRITERIA

We have emphasized earlier that econometric models should be based on economic theory and logic. Therefore, while constricting an econometric model,

you should go by the theoretical appropriateness of including or excluding a variable. In order to have a correctly specified model, a thorough understanding of the theoretical concepts and the related literature is necessary. Also, the model that we fit will only be as good as the data that we have collected. If the data collected does not suffer from, say, multicollinearity or autocorrelation, we are likely to have a more robust model.

As mentioned earlier, the criteria for selecting an appropriate model primarily rests on the theory behind it and the strength of the collected data. Many a time, we observe certain relationship between two variables. Such relationship however may be superficial or spurious. Let us take an example. At a traffic light, cars stop when the signal is red. It does not mean that cars cannot move when there is red light in front of them. It also does not mean that traffic light has some damaging effect on moving cars. The reason is observance of traffic rules. Unless we look into the traffic rules and go by observation only, our reasoning will be wrong. The dependent variable and the independent variable both may be affected by another variable. In such cases the relationship is confounded.

You should note one more issue regarding selection of econometric models. Different test criteria may suggest different models. For example, economic logi suggests that there could two possible econometric models (say, model A and model B) for a particular issue. You may come across a situation such that $\bar{R}^2$ test suggests model A and AIC criterion suggest model B. In such situations you should carry out a number of tests and then only chose the best model.

Adjusted R-squared, Mallows C_p, p-values, etc. may point to different regression equations without much clarity to the econometrician. Thus, we conclude that none of the methods for model selection listed above are adequate by itself. There is no substitute to theoretical understanding of the related literature, accurately collected data, practical understanding of the problem, and common sense while specifying an econometric model. We will discuss further on the model selection criteria in the course BECC 142: Applied Econometrics.

Check Your Progress 1

1) Explain why $\bar{R}^2$ is a better criterion than R^2 in model specification.

.....

.....

.....

.....

.....

.....

2) Explain how the AIC and BIC criteria are applied in selection of econometric models.

.....

.....

.....

.....

.....

3) What precaution you should take while selecting an econometric model?

.....

.....

.....

.....

.....

14.5 LET US SUM UP

Selection of an appropriate econometric model is a difficult task. We have to take into account the economic theory and logic behind the econometric model. There could be many competing models for a particular issue.

There a certain criteria on the basis of which the best econometric model is selected. These criteria could be $\bar{R}^2$, AIC, BIC, and Mallow's C_p. We have described the formulae for these test criteria in the Unit.

14.6 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) In Sub-Section 14.3.1 we have compared between R^2 and $\bar{R}^2$. The $\bar{R}^2$ takes into account the degrees of freedom.

2) You should describe the test statistics used in AIC and BIC criteria (see Section 14.3). The model with lowest value of test statistics is preferred.

3) Go through Section 14.4 and answer.

APPENDIX TABLES

Table A1: Normal Area Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Table A2: Critical Values of Chi-squared Distribution

df\area	0.1	0.05	0.025	0.01	0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.071	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.300
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.195	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672

Table A3: Critical Values of *t* Distribution

Df\p	0.25	0.10	0.05	0.025	0.01	0.005
1	1.0000	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.8165	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.7649	1.6377	2.3534	3.1825	4.5407	5.8409
4	0.7407	1.5332	2.1318	2.7765	3.7470	4.6041
5	0.7267	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.7176	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.7111	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.7064	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.7027	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.6998	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.6974	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.6955	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.6938	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.6924	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.6912	1.3406	1.7531	2.1315	2.6025	2.9467
16	0.6901	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.6892	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.6884	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.6876	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
20	0.6870	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.6864	1.3232	1.7207	2.0796	2.5177	2.8314
22	0.6858	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.6853	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.6849	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.6844	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.6840	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.6837	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.6834	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.6830	1.3114	1.6991	2.0452	2.4620	2.7564
30	0.6828	1.3104	1.6973	2.0423	2.4573	2.7500
inf	0.6745	1.2816	1.6449	1.9600	2.3264	2.5758

Table A4: Critical Values of *F* Distribution

(5% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.014	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.410	5.192	5.050	4.950	4.876	4.818	4.773	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.688	3.581	3.501	3.438	3.388	3.347
9	5.117	4.257	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.136	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.791	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.073	2.840	2.685	2.573	2.488	2.421	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.237
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.266	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.450	2.336	2.249	2.180	2.124	2.077
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.911
inf	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831

Table A4: Critical Values of *F* Distribution (Contd.)

(5% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	243.906	245.950	248.013	249.052	250.095	251.143	252.196	253.253	254.314
2	19.413	19.429	19.446	19.454	19.462	19.471	19.479	19.487	19.496
3	8.745	8.703	8.660	8.639	8.617	8.594	8.572	8.549	8.526
4	5.912	5.858	5.803	5.774	5.746	5.717	5.688	5.658	5.628
5	4.678	4.619	4.558	4.527	4.496	4.464	4.431	4.399	4.365
6	4.000	3.938	3.874	3.842	3.808	3.774	3.740	3.705	3.669
7	3.575	3.511	3.445	3.411	3.376	3.340	3.304	3.267	3.230
8	3.284	3.218	3.150	3.115	3.079	3.043	3.005	2.967	2.928
9	3.073	3.006	2.937	2.901	2.864	2.826	2.787	2.748	2.707
10	2.913	2.845	2.774	2.737	2.700	2.661	2.621	2.580	2.538
11	2.788	2.719	2.646	2.609	2.571	2.531	2.490	2.448	2.405
12	2.687	2.617	2.544	2.506	2.466	2.426	2.384	2.341	2.296
13	2.604	2.533	2.459	2.420	2.380	2.339	2.297	2.252	2.206
14	2.534	2.463	2.388	2.349	2.308	2.266	2.223	2.178	2.131
15	2.475	2.403	2.328	2.288	2.247	2.204	2.160	2.114	2.066
16	2.425	2.352	2.276	2.235	2.194	2.151	2.106	2.059	2.010
17	2.381	2.308	2.230	2.190	2.148	2.104	2.058	2.011	1.960
18	2.342	2.269	2.191	2.150	2.107	2.063	2.017	1.968	1.917
19	2.308	2.234	2.156	2.114	2.071	2.026	1.980	1.930	1.878
20	2.278	2.203	2.124	2.083	2.039	1.994	1.946	1.896	1.843
21	2.250	2.176	2.096	2.054	2.010	1.965	1.917	1.866	1.812
22	2.226	2.151	2.071	2.028	1.984	1.938	1.889	1.838	1.783
23	2.204	2.128	2.048	2.005	1.961	1.914	1.865	1.813	1.757
24	2.183	2.108	2.027	1.984	1.939	1.892	1.842	1.790	1.733
25	2.165	2.089	2.008	1.964	1.919	1.872	1.822	1.768	1.711
26	2.148	2.072	1.990	1.946	1.901	1.853	1.803	1.749	1.691
27	2.132	2.056	1.974	1.930	1.884	1.836	1.785	1.731	1.672
28	2.118	2.041	1.959	1.915	1.869	1.820	1.769	1.714	1.654
29	2.105	2.028	1.945	1.901	1.854	1.806	1.754	1.698	1.638
30	2.092	2.015	1.932	1.887	1.841	1.792	1.740	1.684	1.622
40	2.004	1.925	1.839	1.793	1.744	1.693	1.637	1.577	1.509
60	1.917	1.836	1.748	1.700	1.649	1.594	1.534	1.467	1.389
120	1.834	1.751	1.659	1.608	1.554	1.495	1.429	1.352	1.254
inf	1.752	1.666	1.571	1.517	1.459	1.394	1.318	1.221	1.000

Table A4: Critical Values of *F* Distribution (contd.)

(1% level of significance)

df2/df1	1	2	3	4	5	6	7	8	9	10
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
inf	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321

Table A4: Critical Values of *F* Distribution (contd.)

(1% level of significance)

df2/df1	12	15	20	24	30	40	60	120	INF
1	6106.321	6157.285	6208.730	6234.631	6260.649	6286.782	6313.030	6339.391	6365.864
2	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	9.888	9.722	9.553	9.466	9.379	9.291	9.202	9.112	9.020
6	7.718	7.559	7.396	7.313	7.229	7.143	7.057	6.969	6.880
7	6.469	6.314	6.155	6.074	5.992	5.908	5.824	5.737	5.650
8	5.667	5.515	5.359	5.279	5.198	5.116	5.032	4.946	4.859
9	5.111	4.962	4.808	4.729	4.649	4.567	4.483	4.398	4.311
10	4.706	4.558	4.405	4.327	4.247	4.165	4.082	3.996	3.909
11	4.397	4.251	4.099	4.021	3.941	3.860	3.776	3.690	3.602
12	4.155	4.010	3.858	3.780	3.701	3.619	3.535	3.449	3.361
13	3.960	3.815	3.665	3.587	3.507	3.425	3.341	3.255	3.165
14	3.800	3.656	3.505	3.427	3.348	3.266	3.181	3.094	3.004
15	3.666	3.522	3.372	3.294	3.214	3.132	3.047	2.959	2.868
16	3.553	3.409	3.259	3.181	3.101	3.018	2.933	2.845	2.753
17	3.455	3.312	3.162	3.084	3.003	2.920	2.835	2.746	2.653
18	3.371	3.227	3.077	2.999	2.919	2.835	2.749	2.660	2.566
19	3.297	3.153	3.003	2.925	2.844	2.761	2.674	2.584	2.489
20	3.231	3.088	2.938	2.859	2.778	2.695	2.608	2.517	2.421
21	3.173	3.030	2.880	2.801	2.720	2.636	2.548	2.457	2.360
22	3.121	2.978	2.827	2.749	2.667	2.583	2.495	2.403	2.305
23	3.074	2.931	2.781	2.702	2.620	2.535	2.447	2.354	2.256
24	3.032	2.889	2.738	2.659	2.577	2.492	2.403	2.310	2.211
25	2.993	2.850	2.699	2.620	2.538	2.453	2.364	2.270	2.169
26	2.958	2.815	2.664	2.585	2.503	2.417	2.327	2.233	2.131
27	2.926	2.783	2.632	2.552	2.470	2.384	2.294	2.198	2.097
28	2.896	2.753	2.602	2.522	2.440	2.354	2.263	2.167	2.064
29	2.868	2.726	2.574	2.495	2.412	2.325	2.234	2.138	2.034
30	2.843	2.700	2.549	2.469	2.386	2.299	2.208	2.111	2.006
40	2.665	2.522	2.369	2.288	2.203	2.114	2.019	1.917	1.805
60	2.496	2.352	2.198	2.115	2.028	1.936	1.836	1.726	1.601
120	2.336	2.192	2.035	1.950	1.860	1.763	1.656	1.533	1.381
inf	2.185	2.039	1.878	1.791	1.696	1.592	1.473	1.325	1.000

Table A5: Durbin-Watson d-statistic Level of Significance = 0.05 k= no. of regressors

	k=1		k=2		k=3		k=4	
n	dL	dU	dL	dU	dL	dU	dL	dU
6	0.6102	1.4002						
7	0.6996	1.3564	0.4672	1.8964				
8	0.7629	1.3324	0.5591	1.7771	0.3674	2.2866		
9	0.8243	1.3199	0.6291	1.6993	0.4548	2.1282	0.2957	2.5881
10	0.8791	1.3197	0.6972	1.6413	0.5253	2.0163	0.3760	2.4137
11	0.9273	1.3241	0.7580	1.6044	0.5948	1.9280	0.4441	2.2833
12	0.9708	1.3314	0.8122	1.5794	0.6577	1.8640	0.5120	2.1766
13	1.0097	1.3404	0.8612	1.5621	0.7147	1.8159	0.5745	2.0943
14	1.0450	1.3503	0.9054	1.5507	0.7667	1.7788	0.6321	2.0296
15	1.0770	1.3605	0.9455	1.5432	0.8140	1.7501	0.6852	1.9774
16	1.1062	1.3709	0.9820	1.5386	0.8572	1.7277	0.7340	1.9351
17	1.1330	1.3812	1.0154	1.5361	0.8968	1.7101	0.7790	1.9005
18	1.1576	1.3913	1.0461	1.5353	0.9331	1.6961	0.8204	1.8719
19	1.1804	1.4012	1.0743	1.5355	0.9666	1.6851	0.8588	1.8482
20	1.2015	1.4107	1.1004	1.5367	0.9976	1.6763	0.8943	1.8283
21	1.2212	1.4200	1.1246	1.5385	1.0262	1.6694	0.9272	1.8116
22	1.2395	1.4289	1.1471	1.5408	1.0529	1.6640	0.9578	1.7974
23	1.2567	1.4375	1.1682	1.5435	1.0778	1.6597	0.9864	1.7855
24	1.2728	1.4458	1.1878	1.5464	1.1010	1.6565	1.0131	1.7753
25	1.2879	1.4537	1.2063	1.5495	1.1228	1.6540	1.0381	1.7666
26	1.3022	1.4614	1.2236	1.5528	1.1432	1.6523	1.0616	1.7591
27	1.3157	1.4688	1.2399	1.5562	1.1624	1.6510	1.0836	1.7527
28	1.3284	1.4759	1.2553	1.5596	1.1805	1.6503	1.1044	1.7473
29	1.3405	1.4828	1.2699	1.5631	1.1976	1.6499	1.1241	1.7426
30	1.3520	1.4894	1.2837	1.5666	1.2138	1.6498	1.1426	1.7386
31	1.3630	1.4957	1.2969	1.5701	1.2292	1.6500	1.1602	1.7352
32	1.3734	1.5019	1.3093	1.5736	1.2437	1.6505	1.1769	1.7323
33	1.3834	1.5078	1.3212	1.5770	1.2576	1.6511	1.1927	1.7298
34	1.3929	1.5136	1.3325	1.5805	1.2707	1.6519	1.2078	1.7277
35	1.4019	1.5191	1.3433	1.5838	1.2833	1.6528	1.2221	1.7259
36	1.4107	1.5245	1.3537	1.5872	1.2953	1.6539	1.2358	1.7245
37	1.4190	1.5297	1.3635	1.5904	1.3068	1.6550	1.2489	1.7233
38	1.4270	1.5348	1.3730	1.5937	1.3177	1.6563	1.2614	1.7223
39	1.4347	1.5396	1.3821	1.5969	1.3283	1.6575	1.2734	1.7215
40	1.4421	1.5444	1.3908	1.6000	1.3384	1.6589	1.2848	1.7209
41	1.4493	1.5490	1.3992	1.6031	1.3480	1.6603	1.2958	1.7205

GLOSSARY

Association : It refers to the connection or relationship between variables

Alternative Hypothesis : It is the hypothesis contrary to the null hypothesis. Null hypothesis and alternative hypothesis are mutually exclusive.

Alternative Hypothesis : In hypothesis testing, alternative hypothesis states a condition that is opposite to the null hypothesis. It is expressed as $H_1: \beta_2 \neq 0$, i.e., the slope coefficient is different from zero. It could be positive or negative.

Analysis of Variance (ANOVA) : This is a technique that breaks up the total variability of data into two parts one statistical and the other random.

ANCOVA Model : This is a model which involves both a quantitative and a dummy variable. The form of such a model will be like: $Y_i = \beta_1 + \beta_2 D + \beta_3 X_i + u_i$.

ANOVA Model : This is a regression model containing only a dummy explanatory variable. The functional form of this is like: $Y_i = \beta_1 + \beta_2 D_i + \mu_i$.

Autocorrelation : The Classical Linear Regression Model assumes that the random error terms are not related to each other. In other words, there exists no correlation between the error terms associated with each observation. This assumption is referred as the assumption of no autocorrelation.

Base or Benchmark Category : The dummy variable which takes the value 0 is referred to as the 'base or benchmark category'.

Continuous Random Variable : It refers to a random variable that can take infinite number of values in an interval are called continuous random variables.

Cochrane-Orcutt Procedure : This is a transformation procedure suggested by Cochrane-Orcutt. It is helpful in estimating the value of the correlation coefficient between the error terms. The transformation, enables the application of the OLS method, and yields estimates of parameters which enjoy the BLUE property.

Confidence Interval Approach : In order to test the population parameter, a confidence interval can be constructed about the true but unknown mean. If the population parameter lies within the confidence interval, the null hypothesis is accepted; otherwise it is rejected.

Classical Linear Regression Model : It refers to a linear regression model that establishes a linear relationship between the variables, based on certain specified assumptions.

Chow Test : This test visualizes the presence of structural change that may result in differences in the intercept or the slope coefficient or both. This in referred to as parameter instability. For examining this we perform Chow Test

Causal Relationship : The relationship between the variables where one can figure out the cause and the effect between the two variables.

Confidence Interval : It is the range of values that determines the probability that the value of the parameter lies within the interval.

Chi-square Distribution : Chi-square distribution is the distribution which is the sum of squares of k independent standard normal random variables.

Composite or Two-Sided Hypothesis : In hypothesis testing, a composite hypothesis covers a set of values that are not equal to the given or stated null hypothesis.

Confidence Interval : It refers to the probability that a population parameter falls within the set of critical values taken from the Table.

Discrete Random Variable : It refers to random variables that can assume only countable values.

Distribution Function : Distribution function of a real valued random variable gives a value at any given sample point in the sample space.

Deterministic Component : It represents the systematic component of the regression equation. It is the expected value of the dependent variable for given values of the explanatory variable.

Econometric Model : These are statistical models specifying relationship between relationships between various economic quantities.

Differential Intercept Coefficient : In the ANOVA model $Y_i = \beta_1 + \beta_2 D_i + \mu_i$, since there is no continuous regression line involved, the slope coefficient β_2 actually measures by how much the value of the intercept term differs between the two categories (e.g. male/female) under consideration. For this reason, β_2 is more appropriately called as the 'differential intercept coefficient'.

Dummy Variable Trap : Response to a dummy variable like gender (male/female), caste (general/SC-ST/OBC), etc. are called as categories. Depending on the 'number' of such categories, we must consider including the number of dummy variables in the regression carefully. Usually, this should be 'one less than the number of categories'. Failing to do this will land us in a situation called as the 'dummy variable trap'. This means we will face a situation of multicollinearity with no unique estimates, or efficient estimates, of the parameters. The general rule for introducing the number of dummies is that, if there are *m* attributes or categories, the number of dummy variables introduced should be '*m* – 1'.

Dummy Variables : There are variables which are qualitative in nature. Also known as dummy variables, these variables are referred differently like: indicator variables, binary variables, categorical variables, dichotomous variables.

Durbin *h*-statistic : The Durbin- Watson technique fails to operate when the regression model involves the lagged value of dependent variable as one of the explanatory variables. In such models, the *h* – statistic, also suggested by Durbin, is useful to identify the presence of autocorrelation in the regression model.

Durbin-Watson Test (*d*-statistic) : The test helps detect a first order autocorrelation. The test statistic employed is:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Estimator : A method of arriving at an *estimate* of a parameter.

Estimation of Parameters : This process deals with estimating the values of parameters based on measured empirical data that has a random component.

Estimation : The process of estimating any population parameter.

F-Distribution : It is a right-skewed distribution used for analysis of variance. F-statistic is used for comparing statistical models and to identify the model that best fits the population.

Forecasting : Forecasting is a technique that predicts the future trends by using historical data. The method of forecasting is generally used to extrapolate the parameters such as GDP or unemployment.

Goodness of Fit : An overall goodness of fit that tells us how well the estimated regression line fits the actual Y values. Such a measure is known as the coefficient of determination, denoted by R^2. It is the ratio of explained sum of squares (ESS) to total sum of squares (TSS).

Glejser Test : The Glejser Test is similar to the Park Test. Obtaining e_i from the original model, Glejser suggests regressing the absolute values of e_i, i.e., $|e_i|$ on the X variable expected to be closely associated with the heteroscedastic variance σ_i^2.

Goldfeld-Quandt Test : In this method of testing for heteroscedasticity, we first arrange the observations in increasing order of X_i variable. Next we exclude C observations in the middle of dataset. Thus, $(n - C)/2$ observations in the first part and $(n - C)/2$ observations in the last part constitute two groups. We then proceed to obtain the respective residual sum of squares RSS_1 and RSS_2. The RSS_1 represents the RSS for the regression corresponding to the smaller X_i values and RSS_2 to that of the larger X_i values. We conduct F-test to check for the presence of heteroscedasticity.

Gauss Markov Theorem : Under the assumptions of classical linear regression model, the least squares estimators are Best Linear Unbiased Estimate (BLUE). This means, in the class of all unbiased linear estimators, the OLS estimates have the minimum or least variance.

Hypothesis : It is a tentative statement that we propose to test. It is based on the limited evidence. Hypothesis is formulated on the basis of economic theory or some logic.

Homoscedasticity : A crucial assumption of the Classical Linear Regression Model (CLRM) in that the error term u_i in the population regression function (PRF) is homoscedastic, i.e., they have the same variance σ^2. Such an assumption is referred to as the assumption of homoscedasticity.

Heteroscedasticity : If the variance of u_i is σ_i^2, i.e., it varies from one observation to another, then the situation is referred to as a case of heteroscedasticity.

Interactive Dummy : This is a variable like *DX* in which there is one dummy variable and one quantitative variable. It is considered in the multiplicative form to enable us to see whether the slope coefficients of two groups are same or different. The functional form of this type of regression is $Y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 (D_i X_i) + u_i$.

Jarque-Bera (J-B) Test : This is an asymptotic or large sample test based on OLS residuals in order to test the normality of the error term. Coefficient of skewness: S, i.e., the asymmetry of PDF. Measure of tallness or height of population distribution function: K

For normal distribution S = 0, K = 3

Jarque and Bera constructed J-Statistics given by

$$J_B = \frac{n}{6}\left[S^2 + \frac{(K-3)^2}{4}\right]$$

Linear Regression : In linear regression models the functional form of the relationship between the variables is linear.

Mathematical Model : A description of system using mathematical concepts

Multicollinearity : The classical linear regression model assumes that there is no perfect multicollinearity, implying no exact linear relationship among the explanatory variables, included in multiple regression models.

MWD test : This is the test for the selection of the appropriate functional form for regression as proposed by Mackinnon, White and Davidson. The test is hence known as the MWD Test.

Null Hypothesis : The null hypothesis (also called Strawman hypothesis) states that there is no relationship between the variables. The coefficients are deliberately chosen as zero to find out whether Y is related to X at all. If X really belongs in the model, we would fully expect to reject the zero-null hypothesis H_0 in favour of the alternatives hypothesis H_1 that it is not zero.

Near or imperfect multicollinearity : The case when two or more explanatory variables are not exactly linear this reinforces the fact that collinearity can be high but not perfect.

"High collinearity" refers to the case of "near" or imperfect" or high multicollinearity.

Null Hypothesis : It is the hypothesis that there is no significant difference between specified population, the observed difference is mainly due to sampling or experimental error.

Normal Distribution : It is a very common probability distribution. The curve is bell-shaped and the area under the normal curve is 1.

Ordinary Least Squares Method : Ordinary Least Squares (OLS) is a method for estimation of the unknown parameters in a linear regression model. The OLS method minimizes the sum of the squares of the errors.

Parameters : It is a measurement of any variable. A numerical quantity that characterizes a given population

Prediction : A regression model explains the variation in the dependent variable on the basis of explanatory variables. Given the values of the explanatory variables, we predict the value of the dependent variable. The predicted value is different from the actual value.

Parameter : A quantity or statistical measure for a given population that is fixed. The mean and the variance of a population are population parameters.

p- value : It is the lowest level of significance when the null hypothesis can be rejected.

Power of Test : The power of any test of statistical significance is defined as the probability that it will reject a false null hypothesis. The value of the power of test is given by $(1 - \beta)$.

Population Regression Function (PRF) : A population regression function hypothesizes a theoretical relationship between a dependent variable and a set of independent or explanatory variables. It is a linear function. The function defines how conditional expectation of a variable Y responds to the changes in independent variable X.

Perfect multicollinearity : The case of perfect multicollinearity mainly reflects the situation when the explanatory variables and perfectly correlated with each other implying the coefficient of correlation between the explanatory variables is 1.

Park-Test : If there is heteroscedasticity in a dataset, the heteroscedastic variance σ_i^2 may be systematically related to one or more of the explanatory variables. In such cases, we can regress σ_i^2 on one or more of such *X*- variables. Such an approach, adopted in the Park-test, helps detect the presence of heteroscedasticity.

Random Variable : A variable which takes on values which are numerical outcomes of a random phenomenon.

Regression : A regression analysis is concerned with the study of the relationship the explained or dependent variable and the independent or explanatory variables.

Residual Term : The actual value of Y is obtained by adding the residual term to the estimated value of Y. The residual term is the estimated value of the random error term of the population regression function.

Ridge Regression : The ridge regressions are the method of resolving the problem of multicollinearity. In the ridge regression, the first step is to standardize the variables both dependent and independent by subtracting the respective means and dividing by their standard deviations.

Statistical Inference : It refers to the process of deducing properties of underlying probability distribution of the parameters by analysing data.

Standard Normal Distribution : It refers to a normal distribution with mean 0 and standard deviation 1.

Statistical Inference : It refers to the method of drawing inference about the population parameter on the basis of random sampling.

Statistical Hypothesis: : It is an assumption about a population parameter. This assumption may or may not be true. This statistical hypothesis is either accepted or rejected on the basis of hypothesis testing.

Stochastic Error : The error term represents the influence of those variables that are not included in the regression model. It is evident that even if we try to include all the factors that influence the dependent variable, there exists some intrinsic randomness between the two variables.

Subsidiary or Auxiliary Regressions : When one explanatory variables X is regressed on each of the remaining X variable and the corresponding R^2 is computed. Each of these regressions is referred as subsidiary or auxiliary regression.

***t*- Distribution** : It refers to a continuous probability distribution that is obtained while estimating mean of normally distributed population where sample size is small and population standard deviation is unknown.

Test of significance Approach : The method of inference used to either reject or accept the null hypothesis. This approach makes use of test statistic to make any statistical inference.

Test Statistic : A test statistic is a standardized value that is computed from a sample during the hypothesis testing. On the basis of test statistics one can either reject or accept the null hypothesis.

Type I Error: : In the statistical hypothesis testing, type I error is the incorrect rejection of true null hypothesis. The value is given by alpha level of significance.

Type II Error : The error that occurs when we accept a null hypothesis that is actually false. It is the probability of accepting the null hypothesis when it is false.

Variance Inflation Factor (VIF) : R^2 obtained variables auxiliary regression may not be completely realiable and is not reliable indicator of collinearity. In this method we modify the formula of var (b_2) and (b_3), var $(b_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-R_2^2)}$

White's General Heteroscedasticity Test : This is a method to test the presence of heteroscedasticity in a regression model. In this, the residuals obtained from original regression are squared and regressed on the original variables, their squared values and their cross-products. Additional powers of original *X* variables can be added.

SOME USEFUL BOOKS

Dougherty, C. (2011). *Introduction to Econometrics*, Fourth Edition, Oxford University Press

Gujarati, D. N. and D.C. Porter (2010). *Essentials of Econometrics*, Fourth Edition, McGraw Hill

Kmenta, J. (2008). *Elements of Econometrics*, Second Edition, Khosla Publishing House

Maddala, G.S., and Kajal Lahiri (2012). *Introduction to Econometrics*, Fourth Edition, Wiley

Wooldridge, J. M. (2014). *Introductory Econometrics: A Modern Approach*, Cengage Learning, Fifth Edition