TEERTHANKER MAHAVEER UNIVERSITY
TMU
MORADABAD
तीर्थंकर महावीर विश्वविद्यालय
सम्यक् दर्शन • सम्यक् ज्ञान • सम्यक् चारित्र

Accredited with NAAC A Grade

12-B Status from UGC

Introductory to Econometrics

BAECCC302

CENTRE FOR DISTANCE AND ONLINE EDUCATION

INTRODUCTORY TO ECONOMETRICS (BAECCC302)

REVIEW COMMITTEE

PROGRAMME COORDINATOR

Mr. Namit Bhatnagar
Assistant Professor
Department of Humanities
Centre for Distance and Online Education (CDOE)
Teerthanker Mahaveer University (TMU)

BLOCK PREPARATION

Ms. Charul Verma
Department of Humanities
Centre for Distance and Online Education (CDOE)
Teerthanker Mahaveer University (TMU)

Secretarial Assistance and Composed By:

Mr. Namit Bhatnagar

COPYRIGHT	:	Teerthanker Mahaveer University
EDITION	:	2024 (Restricted Circulation)
PUBLISHED BY	:	Teerthanker Mahaveer University, Moradabad

Module I

Simple Linear Regression Model

This module is divided into two parts. First part deals with the introduction of the discipline econometrics, its nature, scope, methodology and uses. The second part consists of the description of simple linear regression model, its concepts, methods and properties.

1.1: Introduction to Econometrics

The term econometrics was coined in 1926 by Ragnar A. K. Frisch, a Norwegian economist who shared the first Nobel Prize in Economics in 1969 with Jan Tinbergen, an econometrics pioneer. Although many economists had used data and made calculations long before 1926, Frisch felt the significance of a new term associated with the interpretation and use of data in Economics. Today, Econometrics is a broad area of study within economics and the field changes constantly according to the emergence of new tools and techniques.

Econometrics deals with the measurement of economic relationships and it is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms, combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships. The econometric relationships depict the random

behaviour of economic relationships which are generally not considered in economics and mathematical formulations.

1.1.1 Goals/Uses/Scope of Econometrics

Goals/Scope of Econometrics means the importance or usefulness of the science, Econometrics. There are mainly three main goals for the subject. They are,

- Analysis ,
- Policy making, and
- Forecasting

1) Analysis: Testing Economic Theory

Earlier economic theories started with a set of assumptions concerning the behaviour of some individual units (consumers, producers etc.). From these assumptions the economists derived some general conclusions or laws determining the working process of the economic system. Economic theories thus developed in an abstract level were not tested against economic reality. In other words no attempts were made to examine whether the theories explained the actual economic behaviour of individuals.

Econometrics in fact primarily aims at the verification of economic theories. Under such circumstances we can say that the purpose of the research is 'analysis' i.e., to obtain the empirical evidences to test the explanatory power of economic theories and to decide how well they explain the observed behaviour of economic units. Today, any theory, regardless of its elegance in exposition or its sound logical consistency, cannot be established and generally accepted without some empirical testing.

2) Policy Making: Obtaining numerical estimates of the coefficients of economic relationships for policy simulations

In many cases we apply the various econometric techniques in order to obtain reliable estimates of the individual coefficients of economic relationships from which we can evaluate elasticity or other parameters of economic theory. The knowledge of obtaining the numerical value of (For example, the Marginal concepts in Economics, Concept of multiplier, technical coefficients of production etc) coefficients is very important for the formulation of the economic policies of the governments. It helps to compare the effects of various alternative policy decisions.

For example, the decision of the government about devaluating the currency will depend to a great extent on the numerical values of Marginal propensities of imports and exports and as well as the numerical values of price elasticities of imports and exports (e_i and e_x). If the sum of the price elasticities of imports and exports is less than one ($e_i + e_x < 1$) in absolute value, the devaluation will not help in eliminating the deficit in BOP. This shows that how important is the numerical value of the coefficients of economic relationships. Econometrics can provide such numerical estimates and has become an essential tool for the formulation of sound economic policies.

3) Forecasting the future values of economic magnitudes

Forecasting the values of economic variables is essential in framing different economic policies and econometrics will help a lot for such forecasting and policy framing. For example, suppose government is going to frame its poverty policy, under such circumstances it is necessary to know what

the current employment situation is, what will be the level of poverty in the next five years if the Government doesn't take any apt anti poverty programmes. The facts and figures gained through such estimates will help the government to deal with different situations like: If poverty is low in future, government should take appropriate measures to avoid its occurrence. If poverty in the future is high, government should take appropriate measures to reduce it.

Forecasting is thus becoming increasingly important for the regulation of developed economies as well as for the planning of the economic development for the underdeveloped countries.

1.1.2 Economic Theory, Mathematical economics and Econometrics

Econometrics deals with the measurement of economic relationships. It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships. The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods. The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations. It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical

sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

Econometrics is, mainly, statistical techniques applied to economics. Mathematical Economics would also look at applications of other areas of mathematics. For example, Equilibrium Theory uses a lot of Fixed Point Theorems, which rely on ideas from Analysis and Topology.

Mathematical economics is the application of mathematical methods to represent theories and analyze problems in economics. By convention, these applied methods are beyond simple geometry, such as differential and integral calculus, difference and differential equations, matrix algebra, mathematical programming, and other computational methods. Proponents of this approach claim that it allows the formulation of theoretical relationships with rigor, generality, and simplicity.

Mathematics allows economists to form meaningful, testable propositions about wide-ranging and complex subjects which could less easily be expressed informally. Further, the language of mathematics allows economists to make specific, positive claims about controversial or contentious subjects that would be impossible without mathematics. Much of economic theory is currently presented in terms of mathematical economic models, a set of stylized and simplified mathematical relationships asserted to clarify assumptions and implications.

Broad applications include:

- optimization problems (goal equilibrium), whether of a household, business firm, or policy maker
- Static (or equilibrium) analysis in which the economic unit (such as a household) or economic system (such as a market or the economy) is modelled as not changing
- Comparative statics as to a change from one equilibrium to another induced by a change in one or more factors
- Dynamic analysis, tracing changes in an economic system over time, for example from economic growth.

1.1.3 Methodology of Econometrics

How does the econometrician go ahead in analysing an economic theory? There came the role of **methodology in econometrics**, it is in fact a step-by-step procedure. These steps are:

1. Statement of the theory/hypothesis

A theory should have a prediction. In statistics and econometrics, we also speak of **hypothesis**. Hypothesis is an *if-then* Proposition and the theory is in fact a validated hypothesis. One example is about the value of **the Marginal Propensity to Consume** (MPC) proposed by Keynes, $0<MPC<1$. Other examples could be that lower taxes would increase growth, or maybe that it would increase economic inequality, and that introducing a common currency has a positive effect on trade.

2. Specification of Mathematical Model

A model is a simplified representation of a real-world. It should be representative in the sense that it should contain the salient features of the phenomena under study. In general, one of the objectives in modelling is to have a simple model to

explain a complex phenomenon. Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic.

This is where the algebra enters. We need to use mathematical skills to generate an equation. Assume a theory predicting that more schooling increases the wage, ie, a positive relation between years of schooling and wage rate. In economic terms, we say that the return to schooling is positive on wage.

The equation is: $Y= \beta_0+\beta_1 X$

where; Y, the dependent variable as the variable for wage and β_0 is a constant and β_1 is the coefficient of schooling, and X, the Independent variable is a measurement of schooling, i.e. the number of years in school. We also call β_0 as intercept and β_1 as the slope coefficient. Normally, we would expect both β_0 and β_1 to be positive.

3. Specification of Econometric model

An economic model is a set of assumptions that describes the behaviour of an economy, or more generally, a phenomenon. In practice, generally, all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model. Rest of the variables are dumped in a basket called "disturbances" where the disturbances are random variables whose behaviour is unpredictable. This is the main difference between economic modelling and econometric modelling. This is also the main difference between mathematical modelling and statistical modelling. The mathematical modelling is exact in nature, whereas the econometric modelling contains a stochastic term also.

Here, we assume that the mathematical model is correct but we

need to account for the fact that it may not be so. We add an **error term**, *u* to the equation above. It is also called a **random variable or stochastic variable**. It represents other non-quantifiable or unknown factors that affect Y. It also represents errors of measurements that may have entered the data. The econometric equation is:

$$Y= \beta_0+\beta_1X+U$$

The error term, U, is assumed to follow some sort of statistical distribution.

4. Collection of Data

We need data for the variables above. This can be obtained from government statistics agencies and other sources. A lot of data can also be collected on the Internet in these days. But we need to learn the art of finding appropriate data from the ever-increasing loads of data.

Various types of data are used in the estimation of the model.

A) Time series data

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

B) Cross-section data

The cross-section data give information on the variables concerning individual agents (e.g., consumers or produces) at a given point of time. For example, data collected from a sample of consumers and their family budgets showing expenditures on various commodities by each family, as well as information on family income, family composition and other demographic,

social or financial characteristics is an example of cross section data, as we collect all these information at a point o time

C) Panel data:

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

D) Dummy variable data

When the variables are qualitative in nature, then the data is recorded in the form of the indicator function. The values of the variables do not reflect the magnitude of the data. They reflect only the presence/absence of a characteristic. For example, variables like religion, sex, taste, etc. are qualitative variables. The variable `sex' may takes two values – male or female (for the ease of explaining, we are not considering other genders in this example), the variable `taste' takes values-like or dislike etc. Such values are denoted by the dummy variable. For example, these values can be represented as '1' represents male and '0' represents female. Similarly, '1' represents the liking of a particular taste, and '0' represents the disliking of the taste.

5. Estimation of the model

Here, we quantify β_1 and β_2 i.e. we obtain numerical estimates. This is done by statistical technique called **regression analysis**. This provides the empirical content to the theory under consideration.

6. Hypothesis Testing

In this stage of testing the hypothesis first we have to consider the theory and the hypothesis that we explained in earlier stages. The prediction of our theory was that schooling is good for the wage. Does the econometric model support this

hypothesis? What we do here is called **statistical inference (hypothesis testing)**. Technically speaking, to have this positive relationship between years of schooling and wages, the β_2 coefficient of the econometric model should be greater than 0.

7. Forecasting/ prediction

If the hypothesis testing was positive, i.e. the theory was concluded to be correct; we **forecast** the values of the wage by **predicting** the values of education. For example, how much would someone earn for an additional year of schooling? If the X variable is the years of schooling, the β_2 coefficient gives the answer to the question.

8. Use of the model for policy purpose

Lastly, if the theory seems to make sense and the econometric model was not refuted on the basis of the hypothesis test, we can go on to use the theory for policy recommendation.

1.1.4 Types of Econometrics

Econometrics can be classified in to theoretical and applied econometrics.

The **theoretical econometrics** includes the development of appropriate methods for the measurement of economic relationships which are not meant for controlled experiments conducted inside the laboratories. The econometric methods are generally developed for the analysis of non-experimental data.

The **applied econometrics** includes the application of econometric methods to specific branches of econometric theory and problems like demand, supply, production, investment, consumption etc. The applied econometrics

involves the application of the tools of econometric theory for the analysis of the economic phenomenon and forecasting economic behaviour.

1.1.5 Limitations of Econometrics

Like any other subjects, econometrics also not free from limitations. Some of them are;

a) It is concerned only with quantifiable phenomena like prices, production, employment etc. It throws very little light on qualitative problems.

b) All the econometric analysis is based on data availability. The available data may be insufficient and inaccurate.

c) Predictions are made through sampling methods. Therefore, the limitations of the sampling method are also became the limitations of econometrics.

d) The statistical methods used in econometrics are based on certain assumptions, which are not true with economic data.

e) Econometric methods are time consuming, tedious and complex. It requires a sound knowledge of mathematics and statistics.

1.2: Simple Linear Regression Model

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

1.2.1 Linear regression model

The term 'Regression' was introduced by Francis Galton and Galton's Law of Universal regression was confirmed by his Friend, Karl Pearson. The modern interpretation of regression is quiet different from their analysis. By using modern interpretation of regression, we may say that,

'Regression analysis is concerned with the study of the dependence of one variable (dependant variable), on one or more other variables, the explanatory variables (Independent Variable), with a view to estimating and/or predicting the mean or average value of the former in terms of the known or fixed values of the later'

That is, the major objectives of regression analysis are;

- To estimate the mean value of the dependant variable given the value of the independent variables.
- To test the hypothesis suggested by the underlying economic theory about the nature of the dependence.
- To predict or forecast the mean value of the dependant variable, given the values of the independent variables.

That is, through the regression analysis, we are going to estimate the mean value, or average value or expected value of the dependant variable 'Y' based on the known values of the independent variables 'X's. That is we are estimating

$E(Y/X_i)$

ie, conditional expectation of Y given X_i.

Suppose the outcome of any process is denoted by a random variable Y , called as dependent (or study) variable, depends on k independent (or explanatory) variables denoted by X_1, X_2,

X_3.....X_k Suppose the behaviour of Y can be explained by a relationship given by,

$$Y = f(X_1, X_2, ..., Xk, \beta_1, \beta_2, ... \beta_k) + u$$

where f is some well-defined function and $\beta_1, \beta_2, ... \beta_k$ are the parameters which characterize the role and contribution of X_1 , X_2 ,... X_k respectively. The term u reflects the stochastic nature of the relationship between Y and X_1 , X_2 ,..., X_k and indicates that such a relationship is not exact in nature. When u=0, then the relationship is called the mathematical model otherwise the statistical model.

Here we are discussing a 'simple' 'linear' regression model. The term "**model**" is broadly used to represent any phenomenon in a mathematical framework. To explain a simple linear model, two terms 'simple' and 'linear' must be explained first.

The term 'simple' regression means, it is a regression in which the dependant variable is related to a single explanatory variable. It represents a fundamental idea of the regression analysis that 'a model must be as simple as possible. We have the multiple regressions also in which the regressand (Dependent Variable) is related to more than one regressors (Independent Variables). Therefore in a simple regression model there are only two variables;

- One explained variable, and
- One explanatory variable.

For example, in the Keynesian theory of consumption, we are trying to analyse the relationship of consumption with the household income. Here consumption is the regressand and the household income is the regressor. This type of analysis is

called simple regression or two variable regression analysis. This simple form can be expressed as;

$$Y= \beta_0+\beta_1 Xi$$

A model or relationship is termed as 'linear' if it is linear in parameters and 'non-linear', if it is not linear in parameters. In other words, if all the partial derivatives of *Y* with respect to each of the parameters $\beta_1, \beta_2, \ldots\ldots \beta_k$ are independent of the parameters, then the model is called as a **linear model**. If any of the partial derivatives of *Y* with respect to any of the $\beta_1, \beta_2, \ldots\ldots \beta_k$ is not independent of the parameters, the model is called non-linear. Note that the linearity or non-linearity of the model is not described by the linearity or non-linearity of explanatory variables in the model.

For example,

$$y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$$

is a linear model because $\partial y / \partial \beta_i, \ (i = 1, 2, 3)$ are independent of the parameters β_i, (i=1,2,3,...). On the other hand,

$$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X + \varepsilon$$

is a non-linear model because $\partial y / \partial \beta_1 = 2\beta_1 X_1$ depends on β_1, although $\partial y / \partial \beta_2$ and $\partial y / \partial \beta_3$ are independent of any of the β_1, β_2 or β_3.

When the function *f* is linear in parameters, then $Y = f(X_1, X_2, \ldots, X_k, \beta_1, \beta_2, \ldots\ldots \beta_k)+u$ is called a linear model and when the

function f is non-linear in parameters, then it is called a non-linear model. In general, the function f is chosen as

$Y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ... \beta_k) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ to describe a linear model. Since $X_1, X_2, ..., X_k$ are pre-determined variables and Y is the outcome, so both are known. Thus the knowledge of the model depends on the knowledge of the parameters $\beta_1, \beta_2, ... \beta_k$. The statistical linear modelling essentially consists of developing approaches and tools to determine $\beta_1, \beta_2, ... \beta_k$ in the linear mode $Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ given the observations on y and $X_1, X_2, ..., X_k$.

Therefore, simple linear regression model (SLRM) means that,

- There are only two variables; one dependant and one independent and
- The relation between dependant and independent variables are linear in parameters.(may or may not be linear in variables)

Different statistical estimation procedures, e.g., method of maximum likelihood, the principle of least squares, method of moments etc. can be employed to estimate the parameters of the model. The method of maximum likelihood needs further knowledge of the distribution of Y whereas the method of moments and the principle of least squares do not need any knowledge about the distribution of Y. The regression analysis is a tool to determine the values of the parameters given the data on Y and $X_1, X_2, ..., X_k$. Before going in to the process of estimation, it is better to have an idea of some important terms and terminologies such as population regression function, sample regression function, significance of stochastic disturbance term etc. that are frequently used in the analysis of regression models.

1.2.2 Population Regression Function (PRF)

The group of individuals or items under study is known as the population. In statistics, population is the aggregate of facts or objects, animate or inanimate, under study in any statistical investigation. Informally, it means the set of all possible outcomes of an experiment or measurement. We always expect an idealistic situation which is possible only with a population. This is a highly idealistic situation and very rare to occur. Even the distinction between population and sample is relative and such a distinction is necessary in economic studies.

A Population Regression Function (PRF) can be defined as the average value of the dependant variable for a given value of the independent variable. In other words, PRF tries to find out how the average value of the dependant variable varies with the given value of the explanatory variable.

To explain the PRF. We are taken an example of a hypothetical country with a total population of 50 families. Suppose we are interested in studying the relationship between weekly family expenditure (Y) and weekly disposable income (X). That is, we want to predict the (population) mean level of weekly consumption expenditure knowing the family's weekly income. Suppose we divide these 50 families into 10 groups of approximately same income and examine the consumption expenditure of families in each of these income groups. This hypothetical data can be shown as follows.

Table 1.1

X Weekly family income ⇨	10	12	15	18	20	23	25	28	30	35
Y weekly family expenditure ⇩	5 6 7 8 9	5 8 9 10	8 9 10 11 13 15	11 12 13 14 15	11 14 15 16	15 16 17 18 19 23	16 17 18 21 23	16 18 20 22 24 26	20 21 23 24	22 23 24 26 30
Total	35	32	66	65	56	108	95	126	88	125

The above table can be interpreted as follows.

Corresponding to the weekly income 10, there are 5 families whose weekly consumption expenditure range between 5 and 9. Similarly, given X=30, there are 4 families whose weekly consumption expenditure falls between 20 and 24. In other words, each column of the table above gives the conditional distribution of 'Y' conditional upon the given values of X.

From this table we can easily compute conditional probabilities of y, P(Y/X), as follows

For X=10, there are 5 y values, 5,6,7,8 and 9. Therefore, given X= 10, the probability of obtaining any of these consumption expenditure is 1/5. Symbolically,

$$P(Y=7/X=10) = 1/5. \text{ Or}$$

$$P(Y =23/X=30) = ¼ \text{ and so on.}$$

These conditional probabilities can be given in Table 1.2 for the values given in Table 1.1

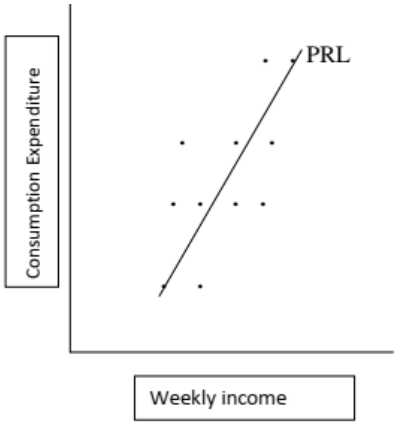
Table 1.2

X ⇨	10	12	15	18	20	23	25	28	30	35
P(y/Xi) ⇩ Conditional probabilities	1/5 1/5 1/5 1/5 1/5	¼ ¼ ¼ ¼	1/6 1/6 1/6 1/6 1/6 1/6	1/5 1/5 1/5 1/5 1/5	¼ ¼ ¼ ¼	1/6 1/6 1/6 1/6 1/6 1/6	1/5 1/5 1/5 1/5 1/5	1/6 1/6 1/6 1/6 1/6 1/6	¼ ¼ ¼ ¼	1/5 1/5 1/5 1/5 1/5
Conditional Means of y	7	8	11	13	14	18	19	21	22	25

From the conditional probability distribution of Y we can compute its mean or average value, known as the, conditional mean or conditional expectation denoted by $E(Y/X=X_i)$ and simply as $E(Y/X_i)$.

When we plot the data of weekly family consumption expenditure at different levels of income on a graph paper we get a scatter diagram as follows (Figure 1.1).

Figure 1.1 Population Regression Line



The Figure 1.1 shows very clearly that consumption expenditure on an average increases as income increases. When we connect the conditional means of Y, we get a straight line with a positive slope. This line is known as the Population Regression Line (PRL).

Geometrically, a population regression curve is simply the locus of the conditional means or expectations of the dependent variable for the fixed values of the independent variables. Therefore it is clear that each conditional mean is a function of X_i. Symbolically,

$$E(Y/X_i) = f(X_i) \text{.....(1)}$$

The above equation is known as the Population Regression Function (PRF). It merely states that the population means of the distribution of Y given X is functionally related to Xi.

Taking our example of consumption function,

$$E(Y/X_i) = \beta_0+\beta_1 X_i \text{.....(2)}$$

The above example is a linear population regression function.

2.2.1 Stochastic Specification of PRF

The PRF states that the mean or average responses of Y varies with X. Consider our consumption expenditure figures, at income level 20, it can be seen as 11, 14, 15 and 16 with a mean 14. The PRF indicates the average only. The deviation from mean of the actual expenditure figures are not explained by the PRF. Therefore, when we take one consumer at random, his consumption expenditure may be greater or less than the mean value. So this can be expressed by the stochastic specification of PRF as;

$$Yi = E(Y/X_i) + u_i$$

Where; u_i = Stochastic error term.

u_i may be defined as an unobservable random variable taking positive or negative values. It is also termed as Stochastic disturbance term.

Therefore, we can explain the stochastic PRF as; the expenditure of a family, given its income level as the sum of two components.

1. $E(Y/X_i)$ - conditional mean expenditure and
2. u_i - random component.

Therefore; our estimated consumption function can be expressed as;

$$Yi = E(Y/X_i) + u_i \text{ or}$$

$$Yi = \beta_0+\beta_1X_i + u_i \text{ .....(3)}$$

The stochastic specified PRF clearly shows that there are other variables besides income that affect consumption expenditure and that an individual family's consumption expenditure cannot be fully explained by the regression model.

<u>The significance of Error term/ Stochastic Disturbance term (u_i</u>

Here we are going to explain the significance of incorporating the term u_i in the econometrics model. Following are the main reasons for its inclusion:

1. Vagueness of theory:- Sometimes we may ignorant or unsure about the other variables affecting Y (consumption) other than income. Under such circumstances, the theory may be incomplete to explain the behaviour of Y. Therefore u_i may be used as a substitute for all the excluded or omitted variables from the model.

2. Unavailability of data:- Even if we know all other variables affecting Y other than X, there may not have quantitative information about these variables. Therefore we may forced to omit some variables from our model despite its great theoretical relevance. Hence u_i may be implied this omitted variables.

3. Core variables versus Peripheral variables:- Apart from the influence of different variables, there may be some variables which jointly influence the model, which is not explicitly in the model. u_i thus tries to explain the combined effect of these variables in the model concerned.

4. Intrinsic randomness in human behaviour:- Even if we succeed in introducing all the relevant variables into the model there may be some intrinsic randomness in the human behaviour. Therefore u_i may be well reflect these intrinsic randomness.

5. Poor proxy variables:- Although the classical regression model assumes that the variables Y and X are measured accurately, in practice, the data may be plagued by errors of measurement. But since data on these variables are not directly observable in practice we may use proxy variables. The disturbance term u_i may in this case also represent the errors of measurement.

6. Principle of parsimony:- By this principle, we would like to keep our regression model as simple as possible. It is done by avoiding some variables from the model. Let u_i represents all the omitted variables.

7. Wrong functional form:- Even if we have theoretically correct variables explaining a phenomena, but unfortunately due to the unavailability of data on these

variables, we do not know the form of the functional relationship between the regressand and the regressors. In two variable models, the functional form of the relationship can often be gained from the scatter diagram but in a multiple regression model it is not easy to determine the appropriate functional form graphically. So the errors of the functional form can also be solved by the inclusion of the random variable u_i.

For all these reasons, stochastic disturbances term u_i assume an extremely crucial role in regression analysis.

1.2.3. The Sample Regression Function (SRF)

Practically, it is not possible to rely on population studies always. Under such circumstances we have to rely on sample studies associated with this we face sampling related problems too. Therefore, our task is to estimate the PRF on the basis of the sample information.

For this, we randomly select some of the Y values corresponding to fixed values of X from the given population. In this way, we have to draw so many samples from the population. Considering our example, we can draw a sample from the given population of income and consumption expenditure as follows (Table 1.3).

The Table 1.3 is a sample of consumption expenditure at different levels of income. Like this we have to draw 'n' different samples from the population.

When we estimate the average value of the dependent variable with the help of a sample it is called stochastic Sample Regression Function (SRF) to estimate the PRF. Actually, to estimate the numerical values of βs in PRF, we have to depend

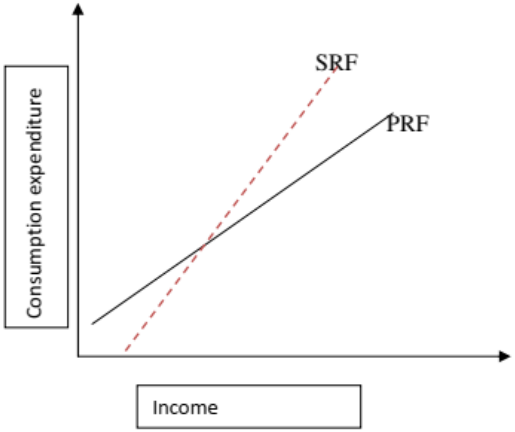
on the whole data. But in practice, we are interested in a sample and with the help of the sample, we are trying to estimate the PRF.

When the plot the sample data on consumption expenditure on a graph paper we have the Figure 1.2

Table 1.3 Sample Data on Consumption Expenditure

Weekly Family Income	Weekly Family consumption Expenditure
10	7
12	8
15	11
18	13
20	15
23	17
25	18
28	20
30	21
35	24

Figure 1.2 Sample Regression Line (SRL)



The SRF can be expressed as;

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \text{.....(4)}$$

Then our objective is to estimate the PRF, $Yi = \beta_0+\beta_1 Xi + u_i$ on the basis of the SRF, $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$.

Here the SRF is the estimator of the PRF. We are using SRF to find PRF. That is, $\hat{\beta}_0$ as the estimator of β_0, $\hat{\beta}_1$ as the estimator of β_1 and $\hat{u}_i$ as the estimator of U_i.

The graphical representation of SRF is termed as sample regression line SRL. To conclude, we can say that the primary objective of regression analysis is to estimate PRF on the basis of the SRF. We may have to select as many samples as possible to reduce the sampling fluctuations, so that it will become more easy to approximate the SRF to the PRF.

So in the linear regression analysis we are trying to estimate the average value of the dependent variable Y for any given values of the independent variable Xi. For this we are estimating β_0 and β_1 with the help of $\hat{\beta}_0$ and $\hat{\beta}_1$.For the estimation of a linear regression model there are mainly two methods;

1. Ordinary Least Square (OLS) method, and
2. Maximum Likelihood (ML) method

The method of OLS is extensively used in regression analysis because of its mathematical simplicity. But both methods gave the same results.

1.2.4 Method of Ordinary Least Squares (OLS Method)

The method of Ordinary Least Squares (OLS) was developed

by a famous German mathematician Carl Frederich Gauss. The OLS method has some very attractive statistical properties under certain assumptions which made it one of the most powerful and popular method of regression analysis. The OLS method can be explained as follows.

Using OLS method through regression analysis we are estimating the PRF:

$$Yi = \beta_0+\beta_1 X_i + u_i$$

From SRF:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

We have to know that;

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Therefore;

$$Y_i = \hat{Y}_i + \hat{u}_i \text{ .....} \quad (5)$$

Where;

$\hat{Y}_i$ = estimated value or conditional mean value of Y_i

From equation (5) we have;

$$\hat{u}_i = Y_i - \hat{Y}_i \text{.....} \quad (6) \text{ or}$$

$$\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \text{.....} \quad (7)$$

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \text{.....} \quad (8)$$

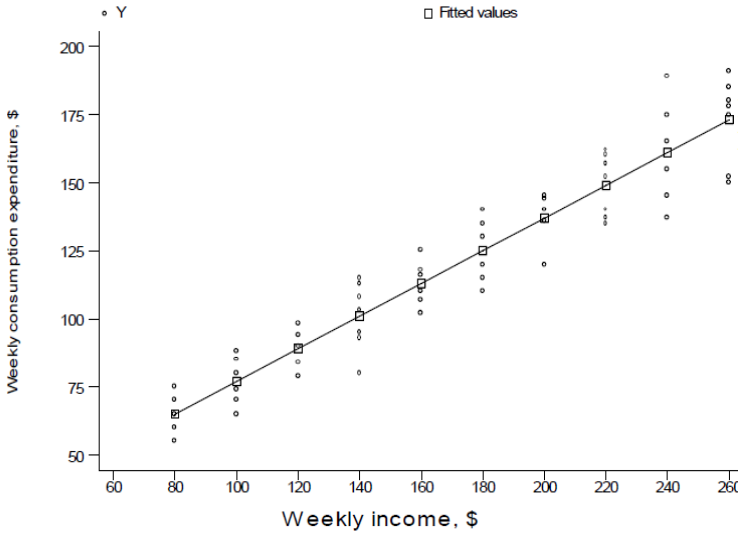
It shows that the residuals ($\hat{u}_i$) are simply the differences between the actual and estimated Y values. For given 'n' pairs of observations on Y and X, we would like to determine the SRF in such a manner that it is as close as possible to the actual Y. To attain this, we may use a criterion that the sum of residuals is as small as possible. That is,

$\Sigma\hat{u}_i = \Sigma (Y_i - \hat{Y}_i)$ is as small as possible.

But it is not a very good criterion because even though the residuals are not scattered evenly from the SRF we gave equal importance to each residuals. To clear this, we have to consider the following starter diagram (Figure 1.3).

When we are considering the scatter diagram (Figure 1.3) some residuals are closer to SRF whereas some others are widely scattered from the SRF. When we are adopting the criterion of minimising $\Sigma\hat{u}_i$ by summing all the $\hat{u}_i$s, so that the algebraic sum of the $\hat{u}_i$ is small or even zero although the $\hat{u}_i$ are widely scattered about the SRF. Even if $\Sigma\hat{u}_i$ is small, we can find a greater difference between actual and estimated Y values. As a result, minimising the sum of squares of errors is not considered to be a very good criterion to estimate PRF using SRF.

Figure 1.3 Scatter diagram of SRF and U_i



We can avoid this problem if we adopt the Least Square criterion. Using this criterion, the SRF can be fixed as;

$\Sigma \hat{u}_i^2 = \Sigma (Y_i - \hat{Y}_i)^2$ is as small as possible

or

$\Sigma \hat{u}_i^2 = \Sigma (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$ is as small as possible.

By squaring the residuals this method gives higher weightage to those residual which are widely scattered about the SRF. So in the OLS method the criteria adopted for fixing SRF is that the sum of squares of the residuals should be minimum to get best estimators.

The OLS principal chooses $\hat{\beta}_o$ and $\hat{\beta}_1$ in such a manner that for a given sample $\Sigma \hat{u}i^2$ is as small as possible. In other words, for a given sample, the method of least squares provides us with unique estimates of $\hat{\beta}_o$ and $\hat{\beta}_1$ that give the smallest possible value of $\Sigma \hat{u}i^2$.

Using OLS method the sum of squared residuals ($\Sigma \hat{u}i^2$) is to be minimised with respect to parameters. For this we are adopting two principles or two steps.

1) Differentiation principle and

2) Minimization principle

That is, we are first differentiating the residuals with respect to parameters. That is we are finding,

$\partial \Sigma \hat{u}i^2 / \partial \hat{\beta}_o$ and $\partial \Sigma \hat{u}i^2 / \partial \hat{\beta}_1$

and then applying the minimization principle;

- First derivative should be equal to zero and

-Second derivative should be greater than zero or positive.

That is the minimization principles are;

$-\partial\Sigma\hat{u}i^2/\partial\ \hat{\beta}_o = 0$ And $\partial\Sigma\hat{u}i^2/\partial\ \hat{\beta}_1 = 0$

And

$-\quad \partial^2\Sigma\hat{u}i^2/(\partial\ \hat{\beta}_o)^2 > 0$ And $\partial^2\Sigma\hat{u}i^2/(\partial\ \hat{\beta}_1)^2 > 0$

By applying these two principles we arrive at two equations, popularly known as normal equations. The derivation of normal equation using differentiation and minimization principles are given below.

$$Y_i = \hat{\beta}_o + \hat{\beta}_1 X_i + \hat{u}_i$$

$$\hat{u}_i = Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i$$

$$\hat{u}_i^2 = (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

Sum of squares,

$$\sum_{i=1}^{n} \hat{u}i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

$$\partial\ \Sigma\hat{u}i^2/\partial\ \hat{\beta}_o = 2\ \sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)$$

$$\partial\ \Sigma\hat{u}i^2/\partial\ \hat{\beta}_o = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} (\hat{\beta}_o + \hat{\beta}_1 X_i)$$

That is,

$$\Sigma Y_i = n\ \hat{\beta}_o + \hat{\beta}_1 \Sigma X_i \quad \text{.....(1)}$$

Similarly,

$\partial \Sigma \hat{u}i^2/\partial \hat{\beta}_1 = 2 \sum_{i=1}^{n}(Y_{i-} \hat{\beta}_o - \hat{\beta}_1 X_i)\ X_i$

$\partial \Sigma \hat{u}i^2/ \hat{\beta}_1 = 0 => \sum_{i=1}^{n}\left(Y_{i-} \hat{\beta}_o - \hat{\beta}_1 X_i\right)\ X_i = 0$

$$\sum_{i=1}^{n} Y_i X_i = \sum_{i=1}^{n}(\hat{\beta}_o + \hat{\beta}_1 X_i)X_i$$

That is,

$$\Sigma Y_i X_i = \hat{\beta}_o \Sigma X_i + \hat{\beta}_1 \Sigma X^2_i \text{ .....(2)}$$

Then the two normal equations are;

$$\Sigma Y_i = n \hat{\beta}_o + \hat{\beta}_1 \Sigma X_i \text{ .....(1)}$$

$$\Sigma Y_i X_i = \hat{\beta}_o \Sigma X_i + \hat{\beta}_1 \Sigma X^2_i \text{ .....(2)}$$

By solving these two normal equations we have;

$$\hat{\beta}_1 = \frac{n\Sigma YiXi - \Sigma Xi\Sigma Yi}{n\Sigma Xi^2 - (\Sigma Xi)^2}$$

$$= \frac{\Sigma\left(X_i - \overline{X}\right)(Y_i - \bar{Y})}{(\Sigma Xi - \overline{X})^2}$$

$$= \frac{\Sigma x_i y_i}{\Sigma {x_i}^2} \quad \text{where, } x_i = X_i - \overline{X} \text{ and } y_i = Y_i - \bar{Y}$$

And,

$$\hat{\beta}_o = \frac{\Sigma Xi^2 \Sigma Yi - \Sigma Xi\, \Sigma YiXi}{n\Sigma Xi^2 - (\Sigma Xi)^2}$$

That is,

$$\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \overline{X}$$

Therefore we have the two OLS estimators of the Simple Linear Regression model as,

$$\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \overline{X} \text{ and}$$

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma {x_i}^2}$$

The estimators thus obtained ($\hat{\beta}_o$ and $\hat{\beta}_1$) are known as the least square estimators or OLS estimators. This OLS estimators posses some statistical properties as;

- The OLS estimators are expressed solely in terms of observable quantities (sample quantities). Therefore they can be easily computed.

- OLS estimators point estimators. That is, given the sample, each estimator will provide only single value of the relevant population parameter

- Once the OLS estimates are obtained from the sample data, the sample regression line can be easily obtained. The regression line thus obtained possess the following properties

* The regression line passes through the sample means of Y and X (X and $\bar{Y}$)

* The mean value of the estimated Y= $\hat{Y}_i$ which is equal to the mean value of the actual Y.

* The mean value of the residuals is zero

* The residuals $\hat{u}_i$ are and correlated with the predicted Y_i

* The residuals are and correlated with X.

1.2.5 The Classical Linear Regression Model

or

The Assumptions underlying the Method of OLS

Like many statistical analyses, ordinary least squares (OLS) regression has its own underlying assumptions. When these assumptions for linear regression are found true, Ordinary Least Squares produces the best estimates. However, if some of these assumptions are not materialised, you might need to employ remedial measures or use other estimation methods to improve the results.

Most of these assumptions pivot around the properties of the error term. Unfortunately, the error term is a population parameter that we never known in advance. Instead, we are using the next best thing that is available—the residuals. Residuals are the sample estimate of the error for each observation.

Assumption 1: The regression model is linear in parameters

This assumption addresses the functional form of the model. In statistics, linearity of a model can be expressed in two ways;

- Linearity in variables and
- Linearity in parameters

Linearity in variables means that the conditional expectation of Y is a linear function of Xi. That is, geometrically the

regression curve in this case is a straight line. In short, the powers of the variables are always one. That is,

$E(Y/Xi) = \beta_0+\beta_1X_i$ is a linear function whereas, $E(Y/Xi) = \beta_0+\beta_1X_i^2$ is not a linear function.

Linearity in parameters means that the conditional expectation of Y is a linear function of parameters, the βs; it may or may not be linear in variables. That is the powers of βs are always one. For example, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$ is a regression model which are linear in parameters whereas, $Y=\beta_0+\beta_1^2X_i$ is not.

Of the two interpretation of linearity, linearity in parameters is relevant for the development of regression theory.

In fact, the defining characteristic of linear regression is this functional form of the *parameters* rather than the ability to model curvature. Linear models can model curvature by including nonlinear *variables* such as polynomials and transforming exponential functions. To satisfy this assumption, the correctly specified model must fit the linear pattern.

Assumption 2: The error term has a population mean of zero

The error term in fact explains the variation in the dependent variable that the independent variables do not explain. Random chance should determine the values of the error term. For our model to be unbiased, the average value of the error term must equal zero. That is, $E(U_i/X_i) = 0$

Suppose the average error is +7. This non-zero average error indicates that our model systematically under-estimates the observed values. Statisticians refer to systematic error like this as bias, and it signifies that our model is inadequate because it is not correct on an average.

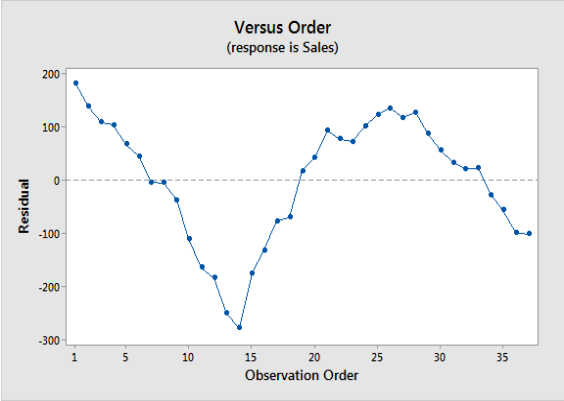
Assumption 3: Observations of the error term are uncorrelated with each other (No Autocorrelation)

This assumption says that no two observations of the error term in a regression model are correlated. That is, one observation of the error term should not predict the next observation. For instance, if the error for one observation is positive and that systematically increases the probability that the following error is positive, that is a positive autocorrelation. If the subsequent error is more likely to have the opposite sign, that is a negative autocorrelation. This problem is known both as serial correlation and autocorrelation. Serial correlation is most likely to occur in time series models. Symbolically no autocorrelation can be expressed as,

$$Cov(U_i, U_j / X_i, X_j) = 0$$

Assess this assumption by graphing the residuals in the order that the data were collected. We want to see randomness in the plot. In the graph for a sales model, there is a cyclical pattern with a positive autocorrelation.

Figure 1.4 Autocorrelation



As it is explained, if we have information that allows to predict the error term for an observation, we must incorporate that information into the model itself. To resolve this issue, we might need to add an independent variable to the model that captures this information. For the sales model shown in Figure 1.4, we need to add variables that explain the cyclical pattern.

Serial correlation reduces the precision of OLS estimates. Analysts can also use time series analysis for time dependent effects.

Assumption 4: The error term has a constant variance (No Heteroscedasticity)

The error variance should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred condition is known as homoscedasticity (same spread/scatter). If the variance changes, we refer to that as heteroscedasticity (different spread/scatter). Homoscedasticity can be expressed symbolically as,

$$Var(U_i/X_i) = \sigma^2$$

Whereas, the hetroscedasticity can be expressed as,

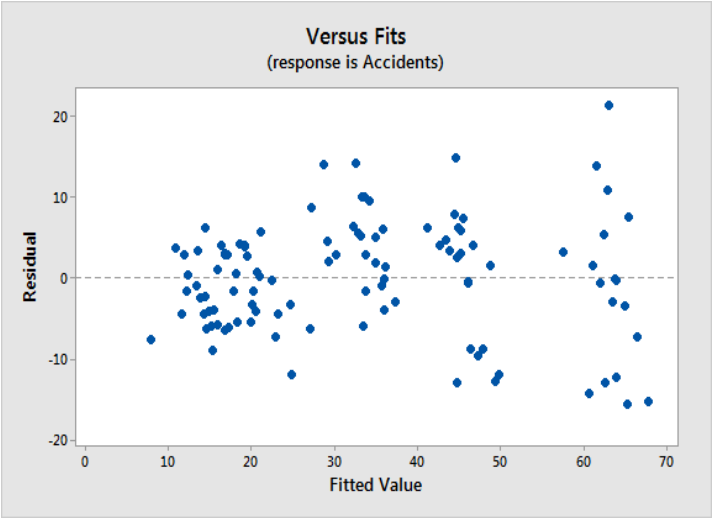
$$Var(U_i/X_i) = \sigma_i^2$$

The easiest way to check this assumption is to create residuals versus estimated value plot. On this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction. In the graph below (Figure 1.5), the spread of the residuals increases as the fitted value increases.

Heteroscedasticity reduces the precision of the estimates in OLS linear regression.

Note: When both assumption 4 (no autocorrelation) and 5 (homoscedasticity) are true, statisticians say that the error term is independent and identically distributed (IID) and

Figure 1.5 Hetroscedasticity



Assumption 5: No independent variable is a perfect linear function of other explanatory variables (No Perfect Multicollinearity)

Perfect correlation occurs when two variables have a Pearson's correlation coefficient of +1 or -1. Under such circumstances, when one of the variable changes, the other variable also change in a fixed proportion.

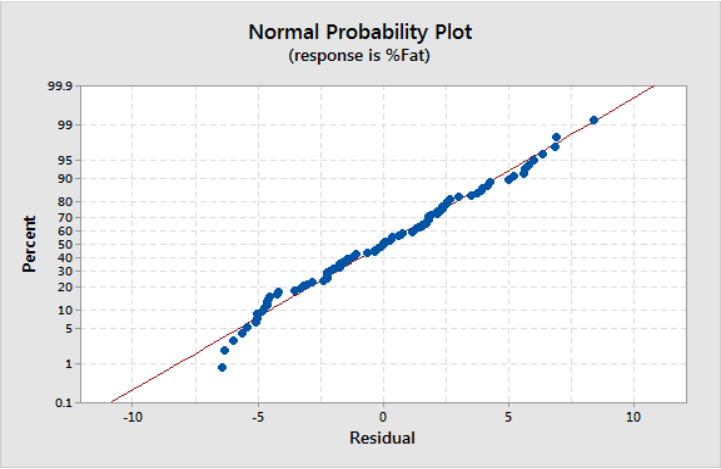
Ordinary Least Squares cannot distinguish one variable from the other when they are perfectly correlated. If these correlations are very high, it can cause problems. Statisticians refer to this condition as multicollinearity, and it reduces the precision of the estimation in OLS linear regression.

Assumption 6: The error term is normally distributed

For estimation of the regression model using OLS does not require the assumption that the error term follows a normal distribution to produce unbiased estimates with the minimum variance. However, satisfying this assumption allows us to perform statistical hypothesis testing and generate reliable confidence intervals and prediction intervals.

The easiest way to determine whether the residuals follow a normal distribution is to assess a normal probability plot. If the residuals follow the straight line on this type of graph, they are normally distributed (Figure 1.6)

Figure 1.6 Normal probability plot



Assumption 7: X values are fixed in repeated sampling

Values taken by the Independent variable X (regressor) are considered to be fixed in repeated sampling. More technically X is assumed to be non-stochastic. When we are estimating PRF based on SRF we may relies on more than one sample. In all these repeated sampling the values of the regressor will

change. This means that our regression analysis is conditional regression analysis, ie, conditional on the given values of the regressor X. That is we are estimating $E(Y/X_i)$.

Assumption 8: Zero covariance between U_i and Xi

This assumption states that the explanatory variables X and the disturbance U are uncorrelated. It is because in our PRF we are assuming that X and U have separate influence on Y. When X and U are correlated it is not possible to assess their individual effects on Y. This zero covariance between X and U can be expressed as;

$$Cov(U_i,X_i) = 0$$

Assumption 9: The number of observations 'n' must be greater than the number of parameters to be estimated

This assumption states that the number of observations associated with X and Y should be greater than the number of parameters (βs) to be estimated. Alternately, the number of observations 'n' must be greater than the number of explanatory variable Xs.

Assumption 10: Variability in X values

This assumption is very important. This assumption states that X values are not identical. That is $X_i \neq \overline{X}$. When $X_i = \overline{X}$, we cannot measure the β_2 and hence the variability in Y. In short, the X values in a given sample must not all be the same. Technically variance of X must be a finite positive number.

$$Var\ (X)>0$$

Simply the assumption states that there must be variability both in X and Y values.

Assumption 11: Rregression model is correctly specified

The CLRM assumed that the model used to test an economic theory is correctly specified. Alternately there is no specification bias or errors in the model used in regression analysis. An econometric investigation begins with the specification of the model underlying the phenomenon of interest. The model specification includes,

-The selection of variables to be included

-The selection of the functional form of the model

-What are the assumptions made about X_i, Y_i and U_i

By doing all these correctly we have valid estimates. The validity of interpreting the estimated regression is highly questionable when the used models are of wrong functional form. Therefore, the correct specification of the economic model is of great importance.

1.2.6 Properties of OLS Estimators

Or

The Gauss-Markov Theorem

Linear regression models have several applications in real life. In econometrics, Ordinary Least Squares (OLS) method is widely used to estimate the parameters of a linear regression model. For the validity of OLS estimates, the following assumptions are made while running linear regression models.

A1. The linear regression model is “linear in parameters.”

A2. There is a random sampling of observations.

A3. The conditional mean should be zero.

A4. There is no multi-collinearity (or perfect collinearity).

A5. Spherical errors: There is homoscedasticity and no auto-correlation

A6: Optional Assumption: Error terms should be normally distributed.

These assumptions are extremely important because violation of any of any of these assumptions would make OLS estimates unreliable and incorrect. Specifically, a violation would result in incorrect signs of OLS estimates, or the variance of OLS estimates would be unreliable, leading to confidence intervals that are too wide or too narrow.

This being said, it is necessary to investigate why OLS estimators and its assumptions gather so much focus. Here, the properties of OLS model are discussed. First, the famous Gauss-Markov Theorem is outlined. Thereafter, a detailed description of the properties of the OLS model is described.

The Gauss-Markov Theorem

The Gauss-Markov Theorem is named after Carl Friedrich Gauss and Andrey Markov.

Let the regression model be:

$$Yi = \beta_0+\beta_1 Xi + u_i$$

Let $\hat{\beta}_o$ is the estimator of β_0 and $\hat{\beta}_1$ is the estimator of β_1. According to the Gauss-Markov Theorem, under the assumptions A_1 to A_5 of the linear regression model, the OLS estimators $\hat{\beta}_o$ and $\hat{\beta}_1$ are the Best Linear Unbiased Estimators (BLUE) of β_0 and β_1. In other words, the OLS estimators $\hat{\beta}_o$ and $\hat{\beta}_1$ have the minimum variance of all linear and unbiased estimators of β_0 and β_1. BLUE summarizes the

properties of OLS regression. These properties of OLS in econometrics are extremely important, thus making OLS estimators one of the strongest and most widely used estimators for unknown parameters. This theorem tells that one should use OLS estimators not only because it is unbiased but also because it has minimum variance among the class of all linear and unbiased estimators.

Properties of OLS Regression Estimators

Property 1: Linear

This property is more concerned with the estimator rather than the original equation that is being estimated. In assumption A_1, the focus was that the linear regression should be "linear in parameters." However, the *linear* property of OLS estimator means that OLS belongs to that class of estimators, which are linear in Y, the dependent variable. Note that OLS estimators are linear only with respect to the dependent variable and not necessarily with respect to the independent variables. The *linear* property of OLS estimators doesn't depend only on assumption A_1 but on all assumptions A_1 to A_5.

Proof:-

We have,

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

$$= \frac{\sum_i x_i (Y_i - \overline{Y})}{\sum_i x_i^2}$$

$$= \frac{\sum_i x_i Y_i}{\sum_i x_i^2} - \frac{\overline{Y} \sum_i x_i}{\sum_i x_i^2}$$

$$= \frac{\sum_i x_i Y_i}{\sum_i x_i^2} \qquad \text{because } \sum_i x_i = 0.$$

Defining the observation weights $k_i = x_i / \sum_i x_i^2$ for i = 1, …, N, we can rewrite the last expression above for as:

$$\hat{\beta}_1 = \sum_i k_i Y_i \qquad \text{where } k_i \equiv \frac{x_i}{\sum_i x_i^2} \qquad (i = 1, ..., N)$$

Note that the formula for $\hat{\beta}_1$ and the definition of the weights k_i imply that is also a linear function of the Y_i such that

$$\hat{\beta}_1 = \sum_i k_i y_i$$

Result: The OLS slope coefficient estimator $\hat{\beta}_1$ is a linear function of the sample values Y_i (i = 1,…,N), where the coefficient of Yi or yi is ki.

Properties of the Weights ki

In order to establish the remaining properties of $\hat{\beta}_1$, it is necessary to know the arithmetic properties of the weights ki. These properties are,

[K1] $\sum_i k_i = 0$, i.e., the weights k_i sum to zero.

$$\sum_i k_i = \sum_i \frac{x_i}{\sum_i x_i^2} = \frac{1}{\sum_i x_i^2} \sum_i x_i = 0, \qquad \text{because } \sum_i x_i = 0.$$

[K2] $\sum_i k_i^2 = \frac{1}{\sum_i x_i^2}$.

$$\sum_i k_i^2 = \sum_i \left(\frac{x_i}{\sum_i x_i^2} \right)^2 = \sum_i \frac{x_i^2}{\left(\sum_i x_i^2\right)^2} = \frac{\left(\sum_i x_i^2\right)}{\left(\sum_i x_i^2\right)^2} = \frac{1}{\sum_i x_i^2}.$$

[K3] $\sum_i k_i x_i = \sum_i k_i X_i$.

$$\begin{aligned} \sum_i k_i x_i &= \sum_i k_i (X_i - \overline{X}) \\ &= \sum_i k_i X_i - \overline{X} \sum_i k_i \\ &= \sum_i k_i X_i \qquad \text{since } \sum_i k_i = 0 \text{ by [K1] above.} \end{aligned}$$

[K4] $\sum_i k_i x_i = 1$.

$$\sum_i k_i x_i = \sum_i \left(\frac{x_i}{\sum_i x_i^2}\right) x_i = \sum_i \frac{x_i^2}{\left(\sum_i x_i^2\right)} = \frac{\left(\sum_i x_i^2\right)}{\left(\sum_i x_i^2\right)} = 1.$$

Implication: $\sum_i k_i X_i = 1.$

Property 2: Unbiasedness

If you look at the regression equation, you will find an error term associated with the regression equation that is estimated. This makes the dependent variable also random. If an estimator uses the dependent variable, then that estimator would also be a random number. Therefore, before describing what unbiasedness is, it is important to mention that unbiasedness property is a property of the estimator and not of any sample.

Unbiasedness is one of the most desirable properties of any estimator. The estimator should ideally be an unbiased estimator of true parameter/population values.

Consider a simple example: Suppose there is a population of size 1000, and you are taking out samples of 50 from this population to estimate the population parameters. Every time you take a sample, it will have the different set of 50 observations and, hence, you would estimate different values of $\hat{\beta}_0$ and $\hat{\beta}_1$. The unbiasedness property of OLS method says that when you take out samples of 50 repeatedly, then after some repeated attempts, you would find that the average of all the $\hat{\beta}_0$ and $\hat{\beta}_1$ from the samples will equal to the actual (or the population) values of β_0 and β_1.

In short,

The **OLS coefficient estimator $\hat{\beta}_1$ is unbiased**, meaning that $E(\hat{\beta}_1) = \beta_1$.

The **OLS coefficient estimator $\hat{\beta}_0$ is unbiased**, meaning that $E(\hat{\beta}_0) = \beta_0$.

Here, 'E' is the expectation operator.

In layman's term, if you take out several samples, keep recording the values of the estimates, and then take an average, you will get very close to the correct population value. If your estimator is biased, then the average will not equal the true parameter value in the population.

Proof:-

We have,

$$\hat{\beta}_1 = \sum_i k_i Y_i$$

Our PRF is, $Y_i = \beta_0 + \beta_1 X_i + u_i$

$$\begin{aligned}
\hat{\beta}_1 &= \sum_i k_i Y_i \\
&= \sum_i k_i (\beta_0 + \beta_1 X_i + u_i) && \text{since } Y_i = \beta_0 + \beta_1 X_i + u_i \text{ by A1} \\
&= \beta_0 \sum_i k_i + \beta_1 \sum_i k_i X_i + \sum_i k_i u_i \\
&= \beta_1 + \sum_i k_i u_i, && \text{since } \sum_i k_i = 0 \text{ and } \sum_i k_i X_i = 1.
\end{aligned}$$

Now take expectations of the above expression for $\hat{\beta}_1$, conditional on the sample values {Xi: i = 1, …, N} of the regressor X. Conditioning on the sample values of the regressor X means that the ki are treated as nonrandom, since the ki are functions only of the Xi.

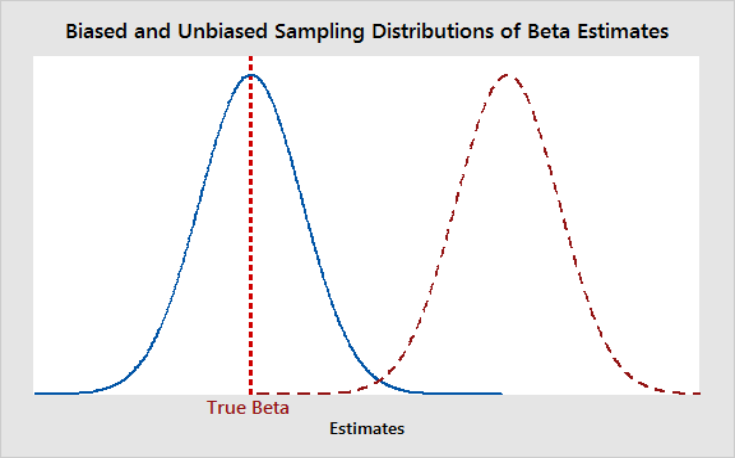
$$\begin{aligned}
E(\hat{\beta}_1) &= E(\beta_1) + E[\sum_i k_i u_i] \\
&= \beta_1 + \sum_i k_i E(u_i | X_i) && \text{since } \beta_1 \text{ is a constant and the } k_i \text{ are nonrandom} \\
&= \beta_1 + \sum_i k_i 0 && \text{since } E(u_i | X_i) = 0 \text{ by assumption A2} \\
&= \beta_1.
\end{aligned}$$

Result: The OLS slope coefficient estimator $\hat{\beta}_1$ is an unbiased estimator of the slope coefficient β_1: that is,

$$E(\hat{\beta}_1) = \beta_1$$

Graphically the property of unbiasedness is depicted in Figure 1.7. The unbiasedness property of OLS in Econometrics is the basic minimum requirement to be satisfied by any estimator. However, it is not sufficient for the reason that most times in real-life applications, we will not have the luxury of taking out repeated samples. In fact, only one sample will be available in most cases.

Figure 1.7 Unbiasedness of OLS Estimators



Property 3: Best: Minimum Variance

First, let us look at what efficient estimators are. The efficient property of any estimator says that the estimator is the *minimum variance unbiased* estimator. Therefore, if we take all the unbiased estimators of the unknown population parameter, the estimator will have the least variance. The estimator that has less variance will have individual data points closer to the mean. As a result, they will be more likely to give better and accurate results than other estimators having higher

variance. In short:

1. If the estimator is unbiased but doesn't have the least variance – it's not the best!
2. If the estimator has the least variance but is biased – it's again not the best!
3. If the estimator is both unbiased and has the least variance – it's the best estimator.

Now, talking about OLS, OLS estimators have the *least variance* among the class of all *linear unbiased* estimators. So, this property of OLS regression is less strict than efficiency property. Efficiency property says least variance among all unbiased estimators, and OLS estimators have the least variance among all linear and unbiased estimators.

Proof:-

The variance of the OLS slope coefficient estimator $\hat{\beta}_1$ is defined as;

$$\mathrm{Var}(\hat{\beta}_1) \equiv E\left\{\left[\hat{\beta}_1 - E(\hat{\beta}_1)\right]^2\right\}.$$

Since $\hat{\beta}_1$ is an unbiased estimator of β_1, $E(\hat{\beta}_1) = \beta_1$. The variance of $\hat{\beta}_1$ can therefore be written as

$$\mathrm{Var}(\hat{\beta}_1) = E\left\{\left[\hat{\beta}_1 - \beta_1\right]^2\right\}.$$

From part (1) of the unbiasedness proofs above, the term $[\hat{\beta}_1 - \beta_1]$, which is called the **sampling error of $\hat{\beta}_1$**, is given by

$$\left[\hat{\beta}_1 - \beta_1\right] = \Sigma_i k_i u_i.$$

The square of the sampling error is therefore

$$\left[\hat{\beta}_1 - \beta_1\right]^2 = (\Sigma_i k_i u_i)^2$$

Since the square of a sum is equal to the sum of the squares plus twice the sum of the cross products,

$$\left[\hat{\beta}_1 - \beta_1\right]^2 = \left(\sum_i k_i u_i\right)^2$$
$$= \sum_{i=1}^{N} k_i^2 u_i^2 + 2\sum_{i<s}\sum_{s=2}^{N} k_i k_s u_i u_s .$$

Now use assumptions A3 and A4 of the classical linear regression model (CLRM):

(A3) $\text{Var}(u_i | X_i) = E(u_i^2 | X_i) = \sigma^2 > 0$ for all i = 1, ..., N;

(A4) $\text{Cov}(u_i, u_s | X_i, X_s) = E(u_i u_s | X_i, X_s) = 0$ for all $i \neq s$.

We take expectations conditional on the sample values of the regressor X:

$$E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] = \sum_{i=1}^{N} k_i^2 E(u_i^2 | X_i) + 2\sum_{i<s}\sum_{s=2}^{N} k_i k_s E(u_i u_s | X_i, X_s)$$
$$= \sum_{i=1}^{N} k_i^2 E(u_i^2 | X_i) \quad \text{since } E(u_i u_s | X_i, X_s) = 0 \text{ for } i \neq s \text{ by (A4)}$$
$$= \sum_{i=1}^{N} k_i^2 \sigma^2 \quad \text{since } E(u_i^2 | X_i) = \sigma^2 \ \forall i \text{ by (A3)}$$
$$= \frac{\sigma^2}{\sum_i x_i^2} \quad \text{since } \sum_i k_i^2 = \frac{1}{\sum_i x_i^2} \text{ by (K2).}$$

Result: The ***variance*** of the OLS slope coefficient estimator $\hat{\beta}_1$ is

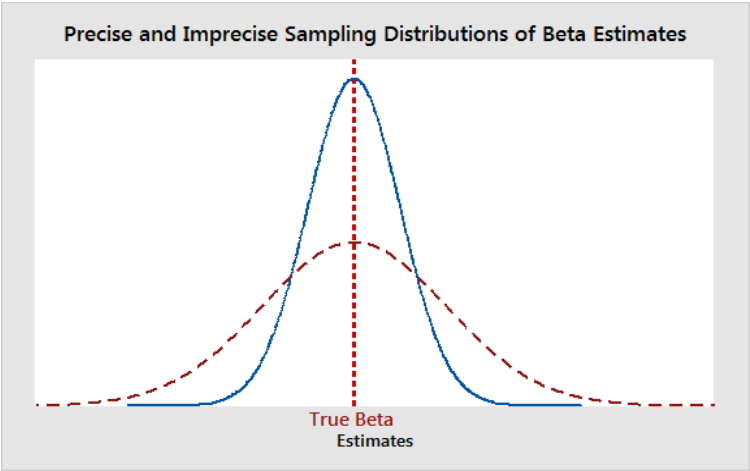
$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i x_i^2} = \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2} = \frac{\sigma^2}{\text{TSS}_X} \quad \text{where } \text{TSS}_X = \sum_i x_i^2$$

The ***standard error*** of $\hat{\beta}_1$ is the square root of the variance: i.e.,

$$\text{se}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)} = \left(\frac{\sigma^2}{\sum_i x_i^2}\right)^{\frac{1}{2}} = \frac{\sigma}{\sqrt{\sum_i x_i^2}} = \frac{\sigma}{\sqrt{\text{TSS}_X}}.$$

Graphically we can show the minimum variance property of OLS estimators as Figure 1.8

Figure 1.8 Minimum Variance of OLS Estimator



The above three properties of OLS model makes OLS estimators BLUE as mentioned in the Gauss-Markov theorem.

It is worth spending time on some other estimators' properties of OLS in econometrics. The properties of OLS described below are asymptotic properties of OLS estimators. So far, finite sample properties of OLS regression were discussed. These properties tried to study the behavior of the OLS estimator under the assumption that you can have several samples and, hence, several estimators of the same unknown population parameter. In short, the properties were that the average of these estimators in different samples should be equal to the true population parameter (unbiasedness), or the average distance to the true parameter value should be the least (efficient). However, in real life, you will often have just one sample. Hence, asymptotic properties of OLS model are discussed, which studies how OLS estimators behave as sample size increases. Keep in mind that sample size should be large.

Property 4: Asymptotic Unbiasedness

This property of OLS says that as the sample size increases, the biasedness of OLS estimators disappears.

Property 5: Consistency

An estimator is said to be consistent if its value approaches the actual, true parameter (population) value as the sample size increases. An estimator is consistent if it satisfies two conditions:

a. It is asymptotically unbiased

b. Its variance converges to 0 as the sample size increases.

Both these hold true for OLS estimators and, hence, they are consistent estimators. For an estimator to be useful, consistency is the minimum basic requirement. Since there may be several such estimators, asymptotic efficiency also is considered. Asymptotic efficiency is the sufficient condition that makes OLS estimators the best estimators.

To conclude, linear regression is important and widely used, and OLS estimation technique is the most prevalent. OLS estimators are BLUE (i.e. they are linear, unbiased and efficient ,have the least variance among the class of all linear and unbiased estimators). Amidst all this, one should not forget the Gauss-Markov Theorem (i.e. the estimators of OLS model are BLUE) holds only if the assumptions of OLS are satisfied. Each assumption that is made while studying OLS adds restrictions to the model, but at the same time, also allows to make stronger statements regarding OLS. So, whenever we are planning to use a linear regression model using OLS, always check for the OLS assumptions. If the OLS assumptions are satisfied, then life becomes simpler, for we can directly use OLS for the best results.

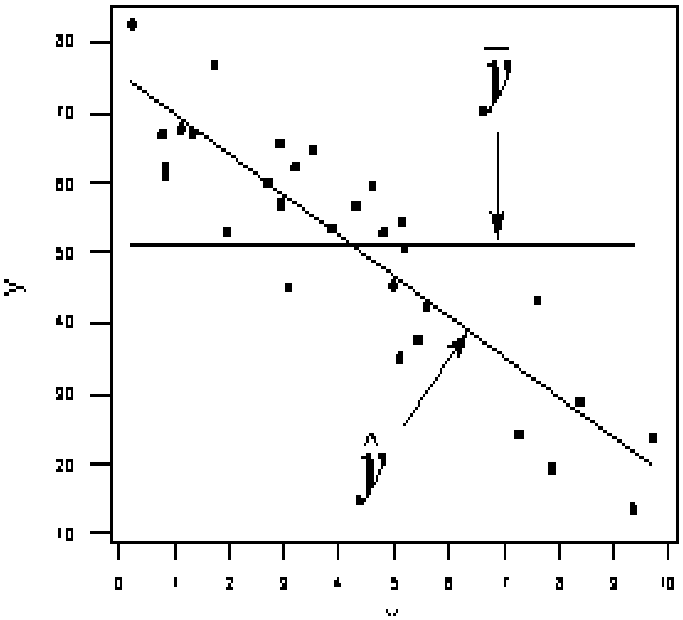
1.2.7 Coefficient of Determination/ Goodness of Fit (r^2)

Here we are considering the "goodness of fit" of the fitted regression line to a set of data. That is, we shall find out how well the sample regression line fits the data. For this we are using the concept 'coefficient of determination'. The **coefficient of determination**, denoted by r^2, is the proportion of the variations in the dependent variable that is predictable from the independent variable. The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event. In other words, this coefficient, which is more commonly known as r^2, assesses how strong the linear relationship is between two variables. The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor.

It is clear that, if all the observations were lie on the regression line, we would obtain a 'perfect fit'. But it is a rare case. This coefficient is commonly known as r^2 and is sometimes referred to as the "goodness of fit." r^2 is simply the square of the sample correlation coefficient (i.e., r) between the observed outcomes and the observed predictor values. This measure is represented as a value between 0.0 and 1.0 ($0 \leq r^2 \leq 1$), where a value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the model fails to accurately model the data at all.

We can show the goodness of fit of a regression line through the graph (Figure 1.9) and from that we can calculate the value of r^2.

Figure 1.9 Goodness of fit of estimated regression line



There are two lines in Figure 1.9, a horizontal line placed at the average response, $\bar{y}$, and a shallow-sloped estimated regression line, $\hat{y}$. From Figure 1.9, the calculation of sum of squares are;

- ***Explained Sum of Squares (ESS)*** quantifies how far the estimated sloped regression line, , is from the horizontal "no relationship line," the sample mean or $\bar{y}$.

  That is, ESS= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

- ***Residual sum of Squares (RSS)*** quantifies how much the data points, yi, vary around the estimated regression line, $\hat{y}$.

That is, RSS = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- ***Total Sum of Squares (TSS)*** quantifies how much the data points, yi, vary around their mean, $\bar{y}$

  That is, TSS = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

The sum of squares can be better illustrated in Figure 1.10.

From these sum of squares, r^2 can be calculated as,

r^2 = ESS/TSS

That is, $r^2 = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

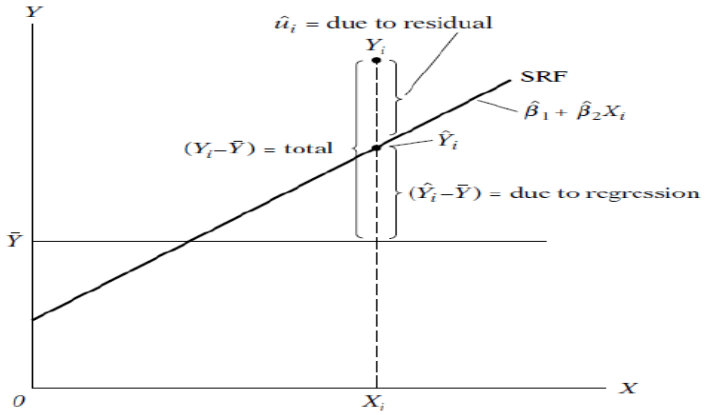
Or in other words,

r^2 = 1- (RSS/TSS) since ESS+RSS= TSS

Therefore;

$$r^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Figure 1.10 Sum of Squares



Here are some basic characteristics of the measure:

- Since r^2 is a proportion, it is always a number between 0 and 1.
- If $r^2 = 1$, all of the data points fall perfectly on the regression line. The predictor x accounts for *all* of the variation in yi
- If $r^2 = 0$, the estimated regression line is perfectly horizontal. The predictor x accounts for *none* of the variation in yi

We've learned the interpretation for the two easy cases — when $r^2 = 0$ or $r^2 = 1$ — but, how do we interpret r^2 when it is some number between 0 and 1. In this situation, the coefficient of determination r^2 can be interpreted as, "$r^2 \times 100$ percent of the variation in Y is explained by the variation in predictor X."

1.2.8 Hypothesis Testing

One important way to make statistical inferences about a population parameter, we use hypothesis testing to make decisions about the parameter's value. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. Hypothesis is in fact an *if-then* proposition. The methodology employed for hypothesis testing depends on the nature of the data used and the objectives to be resolved. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

The null hypothesis (null always indicates zero) is usually a hypothesis of equality between population parameters; e.g., a

null hypothesis may state that the population mean is equal to zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

There are mainly two ways for proceeding with the testing of a hypothesis.

1. The rejection region method

To decide between two competing claims, we can conduct a hypothesis test as follows.

- Express the claim about a specific value for the population parameter of interest as a *null hypothesis*, denoted H0. The null hypothesis needs to be in the form "parameter = some hypothesized value," for example, H0: E(Y) = 255.
- Express the alternative claim as an *alternative hypothesis*, denoted H1. The alternative hypothesis can be in a *lower-tail* form, for example, H1: E(Y) < 255, or an *upper-tail* form, for example, H1: E(Y) > 255, or a *two-tail* form, for example, H1: E(Y) ≠ 255. The alternative hypothesis, also sometimes called the research hypothesis, is what we would like to demonstrate to be the case, and needs to be stated before looking at the data.
- Calculate a *test statistic* based on the assumption that the null hypothesis is true. For testing a univariate population mean, the relevant test statistic is t-statistic.
- Under the assumption that the null hypothesis is true, this test statistic will have a particular probability distribution. For testing a univariate population mean, this t-statistic has a t-distribution with n−1 degrees of freedom. We would

therefore expect it to be "close" to zero (if the null hypothesis is true). Conversely, if it is far from zero, then we might begin to doubt the null hypothesis:

- For an upper-tail test, a t-statistic that is positive and far from zero would then lead us to favor the alternative hypothesis (a t-statistic that was far from zero but negative would favor neither hypothesis and the test would be inconclusive).
- For a lower-tail test, a t-statistic that is negative and far from zero would then lead us to favor the alternative hypothesis (a t-statistic that was far from zero but positive would favor neither hypothesis and the test would be inconclusive).
- For a two-tail test, any t-statistic that is far from zero (positive or negative) would lead us to favor the alternative hypothesis.

- There is always a chance that we might mistakenly reject a null hypothesis when it is actually true. Often, this chance—called the *Level of significance*- will be set at 5%, but more stringent tests (such as in clinical trials of new pharmaceutical drugs) might set this at 1%, while less stringent tests (such as in sociological studies) might set this at 10%. For the sake of argument, we use 5% as a default value for hypothesis tests in this course (unless stated otherwise).
- The significance level dictates the *critical value(s)* for the test, beyond which an observed t-statistic leads to rejection of the null hypothesis in favor of the alternative. This region, which leads to rejection of the null hypothesis, is called the rejection region. For example, for a significance level of 5%:

- For an upper-tail test, the critical value is the 95th percentile of the t-distribution with n−1 degrees of freedom; reject the null in favor of the alternative if the t-statistic is greater than this.
- For a lower-tail test, the critical value is the 5th percentile of the t-distribution with n−1 degrees of freedom; reject the null in favor of the alternative if the t-statistic is less than this.
- For a two-tail test, the two critical values are the 2.5th and the 97.5th percentiles of the t-distribution with n−1 degrees of freedom; reject the null in favor of the alternative if the t-statistic is less than the 2.5th percentile or greater than the 97.5th percentile.

2. The p-value method

An alternative way to conduct a hypothesis test, firstly we assume again that the null hypothesis is true, but then to calculate the probability of observing a t-statistic as extreme as the one observed or even more extreme (in the direction that favors the alternative hypothesis). This is known as the *p-value* (sometimes also called the observed significance level):

- For an upper-tail test, the p-value is the area under the curve of the t-distribution (with n−1 degrees of freedom) to the right of the observed t-statistic.
- For a lower-tail test, the p-value is the area under the curve of the t-distribution (with n−1 degrees of freedom) to the left of the observed t-statistic.
- For a two-tail test, the p-value is the sum of the areas under the curve of the t-distribution (with n−1 degrees of freedom) beyond both the observed t-statistic and the negative of the observed t-statistic.

If the p-value is too "small," then this suggests that it seems unlikely that the null hypothesis could have been true—so we reject it in favor of the alternative. Otherwise, the t-statistic could well have arisen while the null hypothesis held true—so we do not reject it in favor of the alternative. Again, the significance level chosen tells us how small is small: If the p-value is less than the significance level, then reject the null in favor of the alternative; otherwise, do not reject it.

1.2.8.1 't' test

The' t test is usually used to conduct hypothesis tests on the regression coefficients (βs) obtained from simple linear regression. A statistic based on the 't' distribution is used to test the two-sided hypothesis that the true slope, β_1, equals some constant value, $\beta_{1,0}$. The statements for the hypothesis test are expressed as:

$$H_0 : \quad \beta_1 = \beta_{1,0}$$
$$H_1 : \quad \beta_1 \neq \beta_{1,0}$$

The test statistic used for this test is:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the least square estimate of β1, and se($\hat{\beta}_1$) is its standard error. The value of se($\hat{\beta}_1$) can be calculated as follows:

$$se(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

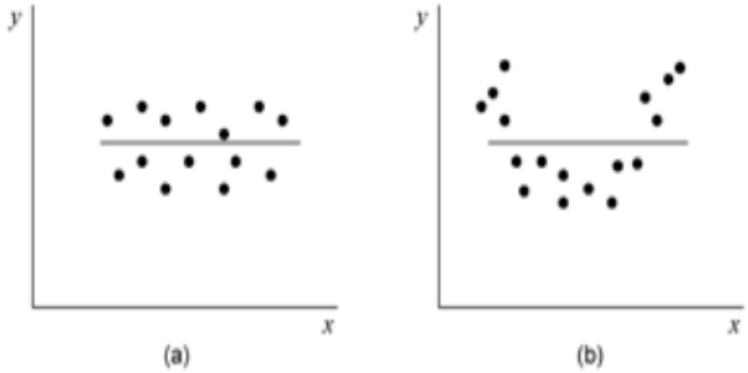
The test statistic, T0 , follows a t distribution with (n−2) degrees of freedom, where n is the total number of observations. The null hypothesis, H0, is accepted if the calculated value of the test statistic is such that:

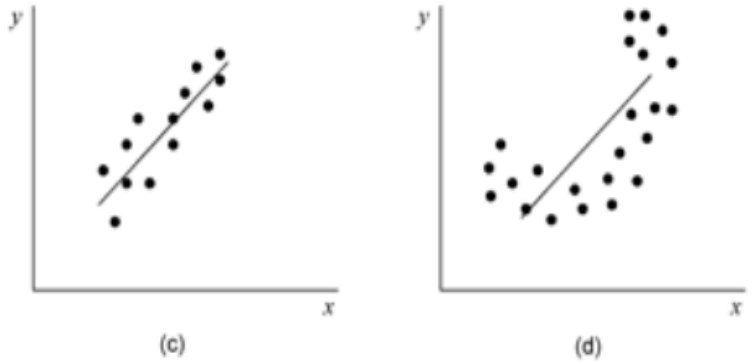
$$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$$

where tα/2,n−2 and −tα/2,n−2 are the critical values for the two-sided hypothesis. tα/2,n−2 is the percentile of the t distribution corresponding to a cumulative probability of (1−α/2) and α is the significance level.

If the value of $\beta_{1,0}$ is zero, then the hypothesis tests for the significance of regression. In other words, the test indicates if the fitted regression model is significant in explaining variations in the observations or if you are trying to impose a regression model when no true relationship exists between x and Y. Failure to reject H0:β1=0 implies that no linear relationship exists between x and Y. This result may be obtained when the scatter plots of against are as shown as Figure 1.11.

Figure 1.11





In Figure 1.11, figure (a) represents the case where no model exits for the observed data. In this case you would be trying to fit a regression model to noise or random variation. (b) represents the case where the true relationship between x and Y is not linear. (c) and (d) represent the case when H0:β1=0 is rejected, implying that a model does exist between x and Y. (c) represents the case where the linear model is sufficient. In the following figure, (d) represents the case where a higher order model may be needed.

A similar procedure can be used to test the hypothesis on the intercept. The test statistic used in this case is:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

where $\hat{\beta}_0$ is the least square estimate of β0, and se($\hat{\beta}_0$) is its standard error which is calculated using:

$$se(\hat{\beta}_0) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

1.2.8.2 F test

***F*-test** is any statistical test in which the test statistic follows an *F*-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact "*F*-tests" mainly arise when the models have been fitted to the data using least squares. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s. Common examples of the use of *F*-tests include the study of the following cases:

- For checking the overall significance of the fitted regression model.
- The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal. This is perhaps the best-known *F*-test, and plays an important role in the analysis of variance (ANOVA).
- The hypothesis that a proposed regression model fits the data well. See Lack-of-fit sum of squares.
- The hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested within each other.

In addition, some statistical procedures, such as Scheffé's method for multiple comparisons adjustment in linear models, also use *F*-tests.

In Simple Linear regression model we are using F test for testing the overall significance of the model. The F-test of

overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables. The overall F-test compares the model that you specify to the model with no independent variables. This type of model is also known as an intercept-only model.

The F-test for testing the overall significance of the model is build on the following two hypotheses:

- The null hypothesis states that the model with no independent variables fits the data as well as your model.
- The alternative hypothesis says that your model fits the data better than the intercept-only model.

In statistical output, you can find the overall F-test in the ANOVA table. An example is below.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	12833.9	4278.0	57.87	0.000
East	1	226.3	226.3	3.06	0.092
South	1	2255.1	2255.1	30.51	0.000
North	1	12330.6	12330.6	166.80	0.000
Error	25	1848.1	73.9		
Total	28	14681.9			

Compare the p-value for the F-test to the pre decided significance level. If the p-value is less than the significance level, the sample data provide sufficient evidence to conclude that our regression model fits the data better than the model with no independent variables.

This finding is good because it means that the independent variables in our model proved to be a better fit in the model.

Generally speaking, if none of the independent variables are statistically significant, the overall F-test is also not statistically significant. Occasionally, the tests can produce conflicting results. Such problems may creep because the F-test of overall significance assesses all of the coefficients jointly whereas the t-test for each coefficient examines them individually. For example, the overall F-test can find that the coefficients are significant *jointly* while the t-tests can fail to find significance *individually (when individual βs are insignificant based on 't'.)*.

These conflicting test results can be hard to understand, but think about it this way. The F-test sums the predictive power of all independent variables and determines that it is unlikely that *all* of the coefficients equal zero. However, it's possible that each variable isn't predictive enough on its own to be statistically significant. In other words, our sample provides sufficient evidence to conclude that our model is significant, but not enough to conclude that any individual variable is significant.

1.2.8.3 Practical versus Theoretical Significance

Theoretical /Statistical significance refer to of how much probabilistically certain you are about an event. If such event is statistically significant, it means that it is highly important in mathematical terms. *E.g.* when you get a very low *p-value* in a test, that's statistically significant because observing such a unusual high (/low) value is very unlikely (given your null hypothesis).

Practical significance refers to the empirical impact that such event has in real life. Obviously, the threshold to define practical significance vary between situations. While statistical

significance relates to whether an effect exists, practical significance refers to the magnitude of the effect.

1.2.9 Method of Maximum Likelihood

We start with the statistical model, which is the Gaussian-noise simple linear regression model, defined as follows:

1. The distribution of X is arbitrary (and perhaps X is even non-random).

2. If X = x, then Y = $\beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients" or "parameters") β_0 and β_1, and some random noise variable ϵ.

3. $\epsilon \sim N(0, \sigma^2)$, and is independent of X.

4. ϵ is independent across observations.

A consequence of these assumptions is that the response variable Y is independent across observations, conditional on the predictor X, i.e., Y1 and Y2 are independent given X1 and X2.

As you'll recall, this is a special case of the simple linear regression model: the first two assumptions are the same, but we are now assuming much more about the noise variable ϵ: it's not just mean zero with constant variance, but it has a particular distribution (Gaussian), and everything we said was uncorrelated before we now strengthen to independence.

Because of these stronger assumptions, the model tells us the conditional probability density function (pdf) of Y for each x, $p(y|X = x; \beta_0, \beta_1, \sigma^2)$. (This notation separates the random variables from the parameters.) Given any data set (x1; y1); (x2; y2); ..... (xn; yn), we can now write down the probability density, under the model, of seeing that data:

$$\prod_{i=1}^{n} p(y_i|x_i;\beta_0,\beta_1,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}}$$

In multiplying together the probabilities like this, we are using the independence of the Yi. When we see the data, we do not known the true parameters, but any guess at them, say (b_0; b_1; s^2), gives us a probability density:

$$\prod_{i=1}^{n} p(y_i|x_i;b_0,b_1,s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(y_i-(b_0+b_1 x_i))^2}{2s^2}}$$

This is the likelihood, a function of the parameter values. It's just as informative, and much more convenient, to work with the log-likelihood,

$$\begin{aligned} L(b_0,b_1,s^2) &= \log \prod_{i=1}^{n} p(y_i|x_i;b_0,b_1,s^2) && (1)\\ &= \sum_{i=1}^{n} \log p(y_i|x_i;b_0,b_1,s^2) && (2)\\ &= -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i-(b_0+b_1 x_i))^2 && (3) \end{aligned}$$

In the method of maximum likelihood, we pick the parameter values which maximize the likelihood, or, equivalently, maximize the log-likelihood. For that, we are using the maximising principle of a function. For that we are firstly differentiating equation (30 with respect to the parameters b_0, b_1 and s^2. After some calculus, this gives us the following estimators:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{c_{XY}}{s_X^2} \qquad (4)$$

$$\hat{\beta}_0 = \overline{y} - b\hat{e}ta_1\overline{x} \quad (5)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad (6)$$

As it is clearly noted, the estimators for the slope and the intercept exactly match the least squares estimators. This is a special property of assuming independent Gaussian noise. Similarly, $\hat{\sigma^2}$ is exactly the in-sample mean squared error.

Module II:
Multiple Regression Analysis

So far, we have seen the concept of simple linear regression where a single independent/predictor variable X was used to model the dependent/response variable Y. Practically, there will be more than one independent variable that influences the response variable. Multiple regression models thus predict how a single response variable Y depends linearly on a number of predictor variables. Examples:

- The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage and a number of other factors.
- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

That is, we use the adjective "simple" to denote that our model has only predictor, and we use the adjective "multiple" to indicate that our model has at least two predictors. The models have similar "LINE" assumptions. The only real difference is that whereas in simple linear regression we think of the distribution of errors at a fixed value of the single predictor, with multiple linear regressions we have to think of the distribution of errors at a fixed set of values for all the predictors. The entire model checking procedures we learned earlier is useful in the multiple linear regression frameworks, although the process becomes more involved since we now have multiple predictors. A population model for a multiple linear regression model that relates a *y*-variable to *k* number of

x-variables is written as,

$$yi=\beta_0+\beta_1 x_{i1},1+\beta_2 x_{i2}+\ldots+\beta_k \ x_{ik}+\epsilon_i.$$

Here we're using "k" for the number of predictor variables, which means we have k+1 regression parameters (the β coefficients).

We assume that the ϵi have a normal distribution with mean 0 and constant variance σ2. These are the same assumptions that we used in simple regression with one x-variable.

The subscript i refers to the ith individual or unit in the population. In the notation for the x-variables, the subscript following *i(1,2...k)* simply denotes which x-variable it is.

The word "linear" in "multiple linear regression" refers to the fact that the model is *linear in the parameters*, $\beta_0,\beta_1,\ldots,\beta_k$. This simply means that each parameter multiplies an x-variable, while the regression function is a sum of these "parameter times x-variable" terms. Each x-variable can be a predictor variable or a transformation of predictor variables (such as the square of a predictor variable or two predictor variables multiplied together). Allowing non-linear transformation of predictor variables like this enables the multiple linear regression models to represent non-linear relationships between the response variable and the predictor variables.

The simplest form of multiple regression models is the three variable regression models which we are going to study in detail in the following session.

2.1 The Three Variable Regression Model

The simplest multiple regression model for two predictor variables is;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.....(1)$$

where

β_0 is the intercept.

$\beta_1$measures the change in y with respect to x_1, holding other factors fixed.

β_2 measures the change in y with respect to x_2, holding other factors fixed.

In short, a partial regression coefficient reflects the effects of one explanatory variable on the mean value of the dependant variable when the values of other explanatory variables included in the model are held constant.

By generalising, the three variable multiple regression model can be written as;

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i.....(2)$$

In the model with two independent variables, the key assumption about how u is related to x_1 and x_2 is;

$$E(u/x_1,x_2) = 0 .....(3)$$

It means that, for any values of x_1 and x_2 in the population, the average of the unobserved factors is equal to zero. Given all other assumptions of classical model, it follows that, on taking the conditional expectation of y on both sides of the equation (2), we have;

$$E(y_i/ x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} .....(4)$$

The equation gives, the conditional mean or expected value of y conditional upon the given or fixed values of the variables x_2 and x_3. Therefore, as in the two variable case, multiple regression analysis is a regression analysis conditional upon

the fixed values of the explanatory variables, and what we obtain is the average or mean value of y or mean response of y for the fixed values of x variables.

2.2 OLS Estimation of Partial Regression Coefficients

To estimate the parameters of the three variable regression model we consider the method of OLS. To find the OLS estimators, let us first write the sample regression function (SRF) corresponding to our PRF.

PRF: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$.....(2)

SRF: $Yi = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}_i$.....(5)

From this we have,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \text{.....(6)}$$

where

$\hat{\beta}_0$ = the estimate of β_0.
$\hat{\beta}_1$ = the estimate of β_1.
$\hat{\beta}_2$ = the estimate of β_2.

But how do we obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? The method of **ordinary least squares** chooses the estimates to minimize the sum of squared residuals. That is, given n observations on y, $x1$, and $x2$, $\{(xi1, xi2, yi): i = 1, 2, \ldots, n\}$, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are chosen simultaneously to make the error sum of squares as small as possible. The error sum of squares is given as;

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \text{.....(7)}$$

The most straight forward procedure to obtain the estimators that will minimise Residual sum of Squares (RSS) is to

differentiate it with respect to the unknowns and set the resulting expression equal to zero and finally solve them simultaneously. This procedure gives the normal equations as;

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$
$$\Sigma Y_i X_{i1} = \hat{\beta}_0 \Sigma X_{i1} + \hat{\beta}_1 \Sigma X_{i1}^2 + \hat{\beta}_2 \Sigma X_{i1} X_{i2}$$
$$\Sigma Y_i X_{i2} = \hat{\beta}_0 \Sigma X_{i2} + \hat{\beta}_1 \Sigma X_{i1} X_{i2} + \hat{\beta}_2 \Sigma X_{i2}^2$$

By simple algebraic manipulations of the preceding equations or simply by solving these normal equations, we obtain,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$
$$\hat{\beta}_1 = \frac{(\Sigma y_i x_{i1})(\Sigma x_{i2}^2) - (\Sigma y_i x_{i2})(\Sigma x_{i1} x_{i2})}{(\Sigma x_{i1}^2)(\Sigma x_{i2})^2 - (\Sigma x_{i1} x_{i2})^2}$$
$$\hat{\beta}_2 = \frac{(\Sigma y_i x_{i2})(\Sigma x_{i1}^2) - (\Sigma y_i x_{i1})(\Sigma x_{i1} x_{i2})}{(\Sigma x_{i1}^2)(\Sigma x_{i2})^2 - (\Sigma x_{i1} x_{i2})^2}$$

Where;

$$y_i = Y_i - \bar{Y}$$
$$x_{i1} = X_{i1} - \bar{X}_1$$
$$x_{i2} = X_{i2} - \bar{X}_2$$

2.3 Multiple coefficient of determination

(R^2 and Adjusted R^2($\bar{R}^2$))

Let R be the multiple correlation coefficient between y , and $x_1,x_2,....x_k$, Then square of multiple correlation coefficient (R^2) is called a coefficient of determination. The value of R^2 commonly describes how well the sample regression line fits the observed data. This is also treated as a measure of **goodness of fit** of the model.

Assuming that the intercept term is present in the model as

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + u_i,\ i = 1, 2, \ldots, n$$

Then,

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= 1 - \frac{SS_{res}}{SS_T} = \frac{SS_{reg}}{SS_T}$$

Where;

SS_{res} = sum of squares due to residuals,

SS_T = total sum of squares

SS reg = sum of squares due to regression.

R^2 measures the explanatory power of the model, which in turn reflects the goodness of fit of the model. It reflects the model adequacy in the sense of how much is the explanatory power of the explanatory variables.

The limits of R^2 are 0 and 1, i.e.,

$0 \leq R^2 \leq 1$.

R^2 = 0 indicates the poorest fit of the model.

R^2 =1 indicates the best fit of the model

R^2 = 0.95 indicates that 95% of the variation in *y* is explained by R^2. In simple words, the model is 95% good. Similarly, any other value of R^2 between 0 and 1 indicates the adequacy of the fitted model.

Adjusted R^2

If more explanatory variables are added to the model, then R^2 increases. In case the variables are irrelevant, then R^2 will still

increase and gives an overly optimistic picture.With a purpose of correction in the overly optimistic picture, adjusted R^2, denoted as $\overline{R}^2$ or adj R^2 is used which is defined as,

$$\overline{R}^2 = 1 - \frac{SS_{res}/(n-k)}{SS_T/(n-1)}$$
$$= 1 - \left(\frac{n-1}{n-k}\right)(1-R^2)$$

We will see later that (n - k) and (n -1) are the degrees of freedom associated with the distributions of SS_{res} and SS_T. Moreover, the quantities $\frac{SS_{res}}{n-k}$ and $\frac{SS_T}{n-1}$ are based on the unbiased estimators of respective variances of e and y in the context of analysis of variance.

The adjusted R^2 will decline if the addition if an extra variable produces too small a reduction in (1- R^2) to compensate for the increase in $\left(\frac{n-1}{n-k}\right)$. Another limitation of adjusted R^2 is that it can be negative also. For example, if k = 3, n =10, $R2$ =0.16, then

$\overline{R}^2 = 1 - \frac{9}{7} \times 0.97 = -0.25 < 0$ which has no interpretation.

Limitations of R^2

1. If the constant term is absent in the model, then R^2 cannot be defined. In such cases, R^2 can be negative. Some ad-hoc measures based on R^2 for regression line through origin have been proposed in the literature.
2. R^2 is sensitive to extreme values, so R^2 lacks robustness.
3. R^2 always increases with an increase in the number of explanatory variables in the model. The main drawback of this property is that even when the irrelevant explanatory

variables are added in the model, R^2 still increases. This indicates that the model is getting better, which is not really correct.

2.4 Testing of Hypothesis - F test

There are several important questions which can be answered through the test of hypothesis concerning the regression coefficients. For example,

1. What is the overall adequacy of the model?
2. Which specific explanatory variables seem to be important? etc.

In order the answer such questions, we first develop the test of hypothesis for a general framework, viz., general linear hypothesis. Then several tests of hypothesis can be derived as its special cases. Here we are considering the **Test of significance of regression (Analysis of variance).**

Under ANOVA,

$$H0: \beta_2=\beta_3=.....\beta_k=0$$

Against the alternative hypothesis

$$H1: \beta_j \neq 0, \text{ for at least one } j = 2,3,..., k$$

This hypothesis determines if there is a linear relationship between y and any set of the explanatory variables X_2, X_3,.....X_k,

Notice that X_1 corresponds to the intercept term in the model and hence $x_{i1}=1$ for all $i= 1, 2,...,n$.

This particular hypothesis explains the goodness of fit. It tells whether β_i has a linear effect or not and are they of any importance. It also tests that X_2, X_3,.....X_k have no influence in the determination of y .

Here;

$\beta_1 = 0$ is excluded because this involves additional implication that the mean level of *y* is zero. Our main concern is to know whether the explanatory variables help to explain the variation in *y* around its mean value or not.

This is an **overall** or **global test of model adequacy**. Rejection of the null hypothesis indicates that at least one of the explanatory variables among X_2, X_3,.....X_k contributes significantly to the model. This is called as **analysis of variance(ANOVA).**

$$\text{Under } H_0 : \beta_2 = \beta_3 = \ldots = \beta_k,$$

$$F = \frac{MS_{reg}}{MS_{res}} \sim F_{k-1,n-k}.$$

The decision rule is to reject at α level of significance whenever

$$F \geq F_{\alpha}(k-1, n-k).$$

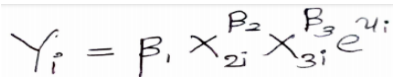
The calculation of *F* -statistic can be summarized in the form of an analysis of variance (ANOVA) table given as follows:

Source of variation	Sum of squares	Degrees of freedom	Mean squares	*F*
Regression	SS_{reg}	$k-1$	$MS_{reg} = SS_{reg}/k-1$	*F*
Error	SS_{res}	$n-k$	$MS_{res} = SS_{res}/(n-k)$	
Total	SS_T	$n-1$		

Rejection of 0 *H* indicates that it is likely that at least one $\beta_i \neq 0$, (i=1,2,....k)

2.5 Restricted least squares

There are occasions where economic theory may suggest that the coefficients in a regression model satisfy some linear equality restrictions. For instance, consider the Cobb-Douglas production function,



Where,

Yi= output

X2= labour input

X3= Capital input

Taking the natural logarithm on both sides we have,

$\text{Ln } Y_i = \text{Ln } \beta_1 + \beta_2 \text{ Ln } X_{2i} + \beta_3 \text{ Ln} X_{3i} + u_i \text{ Ln e}$ ....(2)

$\text{Ln } Y_i = \beta_0 + \beta_2 \text{ Ln } X_{2i} + \beta_3 \text{ Ln} X_{3i} + u_i$ .....(3)

Where,

$\beta_0 = \text{Ln } \beta_1$ and

Ln e = 1

The simplest procedure is to estimate equation (3) is run the OLS. This is called the unrestricted or unconstrained regression.

Now if there are constant returns to scale, economic theory would suggest that,

$\beta_2 + \beta_3 = 1$.....(4)

This is a linear equality restriction.

When estimating equation (3) by considering this linear equality restriction (equation 4) explicitly it is called, restricted

or constrained regression or Restricted Least Square (RLS) when we are using OLS for estimation.

How does one find out if there are constant returns to scale? that is, if the restriction is valid? For this there are two approaches,

1. The 't' test approach and
2. The 'F' test approach

The t-test approach

For applying the t-test approach for checking the validity of the linear restriction, we first estimate the $\hat{\beta}_2$ and $\hat{\beta}_3$ by using the OLS method. Then a test of hypothesis can be conducted by the 't' test equation as,

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - (\beta_2 + \beta_3)}{Se[\hat{\beta}_2 + \hat{\beta}_3]} \quad \text{.....(5)}$$

or

$$t = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{\sqrt{Var(\hat{\beta}_2) + Var(\hat{\beta}_3) + 2\,Cov(\hat{\beta}_2, \hat{\beta}_3)}} \quad \text{..... (6)}$$

Where; $\beta_2 + \beta_3 = 1$

If the t value computed from equation (6) exceeds the critical t value at the chosen level of significance we reject the hypothesis of constant returns to scale otherwise accept it.

The 'F' test approach

The 't' test procedure is a kind of post-mortem examination because we try to find out whether the linear restriction is

satisfied after estimating the unrestricted regression. A direct approach would be to incorporate the restriction into the estimation procedure at the outset. This procedure can be done easily as,

From equation (4) we have ;

$\beta_2 = 1 - \beta_3$ .....(7) or

$\beta_3 = 1 - \beta_2$ .....(8)

Therefore using either of these equalities, we can eliminate one of the coefficients in equation (3). Thus we can write the Cobb-Douglas production function as;

$\text{Ln } Y_i = \beta_0 + (1 - \beta_3) \text{ Ln } X_{2i} + \beta_3 \text{ Ln} X_{3i} + u_i$

$\text{Ln } Y_i = \beta_0 + \text{Ln } X_{2i} - \beta_3 \text{Ln } X_{2i} + \beta_3 \text{ Ln} X_{3i} + u_i$

$\text{Ln } Y_i = \beta_0 + \text{Ln } X_{2i} + \beta_3 (\text{Ln } X_{3i} - \text{Ln} X_{2i}) + u_i$

$(\text{Ln } Y_i - \text{Ln} X_{2i}) = \beta_0 + \beta_3 (\text{Ln } X_{3i} - \text{Ln} X_{2i}) + u_i$.....(9) or

$\text{Ln } (Y_i / X_{2i}) = \beta_0 + \beta_3 (\text{Ln } X_{3i} / X_{2i}) + u_i$.....(10)

Where,

Y_i / X_{2i} = output-labour ratio &

X_{3i} / X_{2i} = capital-labour ratio

These two quantities have great economic importance. The transformed model of Cobb-Douglas production function incorporates the linearity restriction. This procedure will guarantee that the sum of the estimated coefficients of the two inputs will be equal to 1(ie, $\beta_2 + \beta_3 = 1$). Once we estimate from equation (9) by using OLS or from equation (10) β_2 can be easily estimated from the relation (7). This procedure outlined in equation (9) or equation (10) is known as Restricted Least Squares. This procedure can be generalized to models

containing any number of explanatory variables and more than one linear equality restriction.

How do we compare the unrestricted and restricted least square regression? In other words, how do we know that the restriction is valid? By applying 'F' test we can do this. For this,

Let,

$\Sigma\hat{u}^2_{UR}$ = Residual Sum of Squares (RSS) of unrestricted regression (3)

$\Sigma\hat{u}^2_{R}$ = RSS of restricted regression (9)

m = number of linear restrictions

k = number of parameters in the unrestricted regression

n = number of observations

Then,

$$F = \frac{RSS_R - RSS_{UR} / m}{RSS_{UR} / n-K}$$

Or

$$F = \frac{\Sigma \hat{u}^2_R - \Sigma \hat{u}^2_{UR} / m}{\Sigma \hat{u}^2_{UR} / n-K}$$

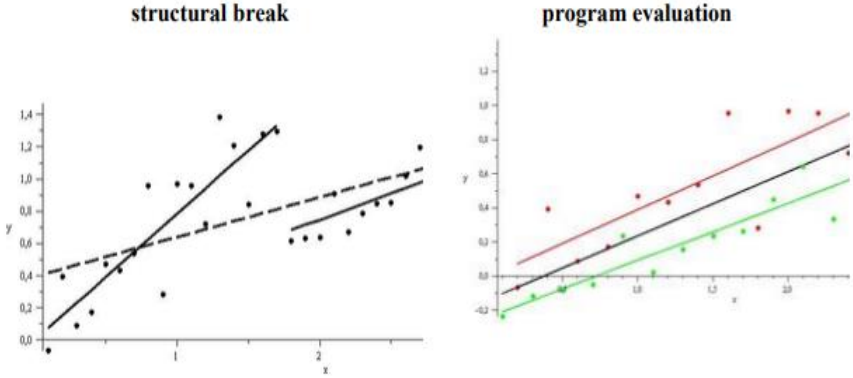
Follows the 'F' distribution with m, (n-k) degrees of freedom. If computed 'F' exceeds critical 'F' we reject the hypothesis of validity of restriction and otherwise accept it.

2.6 Chow test

The Chow test is a statistical and econometric test used to check whether the coefficients in two linear regressions on

different data sets are equal. The Chow test was invented by economist Gregory Chow. In econometrics, the Chow test is most commonly used in time series analysis to test for the presence of a structural break. In program evaluation, the Chow test is often used to determine whether the independent variables have different impacts on different subgroups of the population. These two are shown in Figure 2.1

Figure 2.1 Chow Test



At $x = 1.7$ there is a structural break, regression on the subintervals $[0,1.7]$ and $[1.7,4]$ delivers a better modelling than the combined regression(dashed) over the whole interval.

Comparison of 2 different programs (red, green) existing in a common data set, separate regressions for both programs deliver a better modelling than a combined regression (black).

Suppose that we model our data as,

$$y_t = a + bx_{1t} + cx_{2t} + \varepsilon$$

If we split our data into two groups, then we have,

$$y_t = a_1 + b_1 x_{1t} + c_1 x_{2t} + \varepsilon.$$

and

$$y_t = a_2 + b_2 x_{1t} + c_2 x_{2t} + \varepsilon.$$

The null hypothesis of the Chow test asserts that ,

a1 = a2, b1 = b2, and c1 = c2.

Let,

- SC be the Sum of Squared Residuals from the combined data,
- S1 be the Sum of Squared Residuals from the first group, and
- S2 be the Sum of Squared Residuals from the second group.
- N1 and N2 are the number of observations in each group and k is the total number of parameters (in this case, 3).

Then the Chow test statistic is,

$$\frac{(S_C - (S_1 + S_2))/(k)}{(S_1 + S_2)/(N_1 + N_2 - 2k)}$$

The test statistic follows the F distribution with k and N1 + N2 − 2k degrees of freedom.

2.7 Matrix approach to Multiple Regressions

Using the matrix approach to multiple regressions we are explaining mainly three concepts in matrix notation. They are;

- General k- variable regression model
- Assumptions of CLRM
- Estimation of Multiple Regression Model
- BLUE properties of OLS estimators in the case of multiple regressions.

2.7.1 General k variable regression model

The general multiple linear regression model has k = K−1 regressors; its PRF is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

Here, total number of regression coefficients = K and the number of *slope* coefficients = k = K − 1.

the *n*-sets of observations are also assumed to follow the same model. Thus they satisfy,

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n. \end{aligned}$$

These *n* equations can be written as;

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

In general, the model with *k* explanatory variables can be expressed as;

$$y = X\beta + \varepsilon$$

where $y = (y_1, y_2, \ldots, y_n)$ is a $n \times 1$ vector of *n* observation on study variable.

$y = X\beta + \varepsilon$ is the matrix notation of general k variable multiple regression model where;

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

is a $n \times k$ matrix of n observations on each of the k explanatory variables, $\beta = (\beta_1, \beta_2, ..., \beta_k)$ is a $k \times 1$

vector of regression coefficients and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is a $n \times 1$ vector of random error components or disturbance term.

If intercept term is present, take first column of X to be $(1,1,\ldots,1)$

2.7.2 Assumptions of CLRM in Matrix notation

Some assumptions are needed in the model $y = X\beta + \epsilon$ for drawing the statistical inferences. The following assumptions are made:

(i) $E(\varepsilon) = 0$

(ii) $E(\varepsilon\varepsilon') = \sigma^2 I_n$

(iii) $Rank(X) = k$

(iv) X is a non-stochastic matrix

(v) $\varepsilon \sim N(0, \sigma^2 I_n)$.

These assumptions are used to study the statistical properties of the estimator of regression coefficients. The following assumption is required to study, particularly the large sample properties of the estimators.

(vi) $\lim_{n\to\infty}\left(\frac{X'X}{n}\right) = \Delta$ exists and is a non-stochastic and non-singular matrix (with finite elements).

The explanatory variables can also be stochastic in some cases.

We assume that X is non-stochastic unless stated separately. We consider the problems of estimation and testing of hypothesis on regression coefficient vector under the stated assumption.

2.7.3 Estimation of Multiple regression Model (OLS)

Let B be the set of all possible vectors β. If there is no further information, the B is k -dimensional real Euclidean space. The object is to find a vector $b' = (b_1, b_2, ...b_k)$ from B that minimizes the sum of squared deviations of ϵi 's i.e.,

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

for given y and X. A minimum will always exist as $S(\beta)$ is a real-valued, convex and differentiable function. Write

$$S(\beta) = y'y + \beta'X'X\beta - 2\beta'X'y$$

Differentiate $S(\beta)$ with respect to β

$$\frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y$$

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2X'X \quad \text{(atleast non-negative definite).}$$

The normal equation is

$$\frac{\partial S(\beta)}{\partial \beta} = 0$$

$$\Rightarrow X'Xb = X'y$$

where the following result is used:

Result: If $f(z) = Z'AZ$ is a quadratic form, Z is a $m \times 1$ vector and A is any $m \times m$ symmetric matrix

Then,

$$\frac{\partial}{\partial z}F(z) = 2Az\ .$$

Since it is assumed that rank (X) = k (full rank), then $X'X$ is a positive definite and unique solution of the normal equation is,

$$b = (X'X)^{-1}X'y$$

which is termed as **ordinary least squares estimator** (OLSE) of β.

Since $\frac{\partial^2 S(\beta)}{\partial \beta^2}$ is at least non-negative definite, so b minimize $S(\beta)$ **.**

2.7.4 Properties of OLS estimators

(i) Estimation error:

The estimation error of b is'

$$\begin{aligned} b-\beta &= (X'X)^{-1}X'y-\beta \\ &= (X'X)^{-1}X'(X\beta+\varepsilon)-\beta \\ &= (X'X)^{-1}X'\varepsilon \end{aligned}$$

(ii) **Bias**

Since X is assumed to be nonstochastic and $E(\epsilon) = 0$

$$\begin{aligned} E(b-\beta) &= (X'X)^{-1}X'E(\varepsilon) \\ &= 0. \end{aligned}$$

Thus OLS estimator is an unbiased estimator of β.

(iii) **Covariance matrix**

The covariance matrix of b is

$$\begin{aligned} V(b) &= E(b-\beta)(b-\beta)' \\ &= E\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\right] \end{aligned}$$

$$= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'IX(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}.$$

(iv) Variance

The variance of *b* can be obtained as the sum of variances of all $b_1, b_2 \ldots b_k$ which is the trace of covariance matrix of *b* . Thus,

$$Var(b) = tr\left[V(b)\right]$$
$$= \sum_{i=1}^{k} E(b_i - \beta_i)^2$$
$$= \sum_{i=1}^{k} Var(b_i).$$

(v) Gauss-Markov Theorem:

The ordinary least squares estimator (OLSE) is the best linear unbiased estimator (BLUE) of β.

Proof: The OLSE of β is'

$$b = (X'X)^{-1}X'y$$

which is a linear function of *y* . Consider the arbitrary linear estimator $b^* = a'y$ of linear parametric function $\ell'\beta$ where the elements of *a* are arbitrary constants. Then for *b**

$$E(b^*) = E(a'y) = a'X\beta$$

and so *b** is an unbiased estimator of $\ell'\beta$ when

$$E(b^*) = a'X\beta = \ell'\beta$$
$$\Rightarrow a'X = \ell'.$$

Since we wish to consider only those estimators that are linear and unbiased, so we restrict ourselves to those estimators for which $a'X = \ell'$

Further,

$$Var(a'y) = a'Var(y)a = \sigma^2 a'a$$
$$Var(\ell'b) = \ell'Var(b)\ell$$
$$= \sigma^2 a'X(X'X)^{-1}X'a.$$

Consider,

$$Var(a'y) - Var(\ell'b) = \sigma^2\left[a'a - a'X(X'X)^{-1}X'a\right]$$
$$= \sigma^2 a'\left[I - X(X'X)^{-1}X'\right]a$$
$$= \sigma^2 a'(I - H)a.$$

Since (I -H) is a positive semi-definite matrix, so

$$Var(a'y) - Var(\ell'b) \geq 0$$

This reveals that if b^* is any linear unbiased estimator then its variance must be no smaller than that of b.

Consequently b is the best linear unbiased estimator, where 'best' refers to the fact that b is efficient within the class of linear and unbiased estimators.

Module III

Econometric Problems

The simplest econometric model is the ordinary least square model (OLS). This model minimizes the sum of squared errors (deviation between actual values and estimated values of the dependent variable). The classical linear regression model (CLRM) is buit upon some important assumptions.

By relaxing these assumptions of CLRM, we are confronted with some econometric problems. Major econometric problems arises when we relax these assumptions of CLRM are;

1. Hetroscedasticity
2. Autocorrelation and
3. Multicollinearity

3.1. Hetroscedasticity

The classical linear regression model is that the disturbances *ui* appearing in the population regression function are homoscedastic; that is, they all have the same variance. In this lesson we examine the validity of this assumption and find out what happens if this assumption is not fulfilled. We seek answers to the following questions:

- *What is the nature of heteroscedasticity?*
- *What are its consequences?*
- *How can we detect it?*
- *What are the remedial measures?*

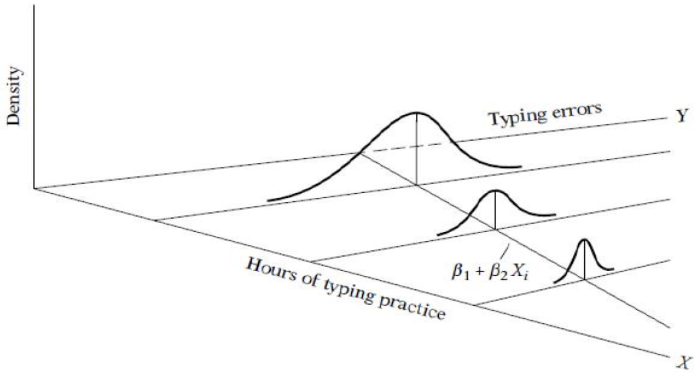
3.1.1 Nature of Heteroscedasticity

Where the conditional variance of the Y population varies with X, this situation is known appropriately as heteroscedasticity or unequal spread or variance. That is,

$$E(u_i^2) = \sigma_i^2$$

We can illustrate the problem of Hetroscedasticity as in Figure 3.1.

Figure 3.1 Hetroscedasticity



3.1.2 Reasons for Hetroscedasticity

Various reasons for the origin of Hetroscedasticity are;

1. In an error learning model, as people learn, their error of behaviour become smaller over time.
2. As income grows, people have more discretionary income & hence more scope for choice about the disposition of their income.
3. As data collecting techniques increases σ_i^2 is likely to decrease.
4. If can also arise as a result of the presence of collinear.

5. If there is skewness in the distribution of one or more regressors included in the model, there is chances of hetroscedasticity.
6. Incorrect data transformation.
7. Incorrect functional form.

3.1.3 Consequences of Heteroscedasticity

In the presence of ***Heteroscedasticity***, we can estimate our regression model and find out the parameters of the model as;

$$E(u_i^2) = \sigma_i^2$$

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

By using the SRF,

$$Yi = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$$

Applying the usual formula the OLS estimator $\hat{\beta}_2$ is

$$\hat{\beta}_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$= \frac{n\Sigma x_i y_i}{n\Sigma X_i} - \frac{\Sigma x_i y_i}{(\Sigma X_i)^2}$$

$$\therefore \quad Var\hat{\beta}_2 = \frac{\sigma_i^2}{\sum x_i^2}$$

Under CLRM, these OLS estimators are BLUE. Now with Heteroscedasticity, the consequences are;

1. OLS estimators are still linear
2. OLS estimators are still unbiased

3. But they no longer have minimum variance. That is, they are no longer efficient. In short, OLS estimators are no longer BLUE in small as well as in large samples.

4. The usual formula to estimate variances of OLS estimators are generally biased. The usual formula is,

$$\therefore \quad Var\hat{\beta}_2 = \frac{\sigma_i^2}{\sum x_i^2}$$

Because of heteroscedasticity we cannot use this formula.

Instead,

$\text{Var}(\hat{\beta}_2) = \frac{\Sigma xi^2 \sigma i^2}{(\Sigma xi^2)^2}$ is used.

A positive bias occurs if OLS overestimates the true variance of estimator and the negative bias occurs if OLS underestimate the true variance of estimator.

5. The bias arises from the fact that the conventional estimator of true σ^2 is no longer and unbiased estimator of σi^2

6. As a result the usual confidence intervals and hypothesis tests based on t and F distributions are unreliable. If conventional testing procedures are employed there is a possibility of drawing wrong conclusions

In short, in the presence of Heteroscedasticity OLS estimators are no longer BLUE. So we rely on other methods like Generalized Least Square (GLS) for estimation. Similarly, ordinary testing of hypothesis is not reliable raising the possibility of drawing wrong conclusions. Therefore it is essential to detect and solve the problem of Heteroscedasticity before estimation.

3.1.4 Detection of Heteroscedasticity

It is noted that there are no hard and fast rules for detecting Heteroscedasticity and we have only a few rules of thumb. But this situation is inevitable because σi^2 can be known only if we have the entire Y population corresponding to the chosen X's. But such data is a rare case in most economic investigations. Therefore in most cases, involving economic investigations Heteroscedasticity may be a matter of intuition, educated guess work, prior empirical experience, or sheer speculation.

Let us examine some of the informal and formal methods of detecting Heteroscedasticity. Most of these methods are based on the examination of the OLS residuals $\hat{u}_i$ since they are the one we observe, and by hoping they are the good estimates of disturbances u_i.

Informal Methods

1. Nature of Problem: - Very often nature of the problem under consideration suggests whether Heteroscedasticity is likely to be encountered. Based on the past studies, one can analyse the nature of hetroscedasticity in the surveys. Now one generally assumes that in similar surveys one can expect unequal variances among the disturbances. As a matter of fact, in cross-sectional data involving heterogeneous units, hetroscedasticity may be the rule rather than exception.

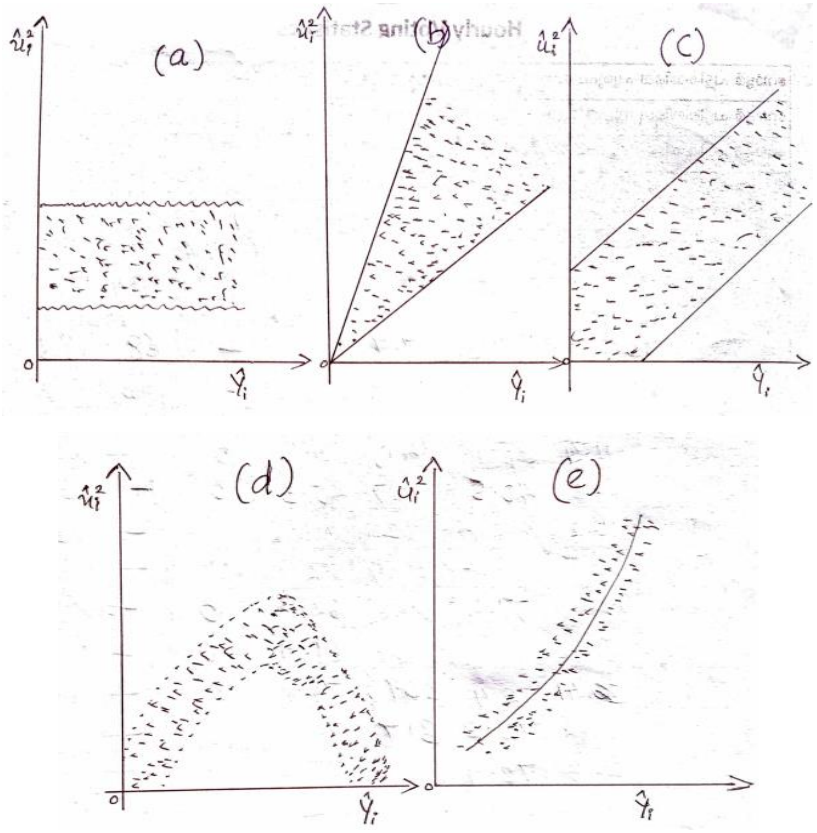
2. **Graphical Method**: - If there is no empirical information about the nature of Heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no Heteroscedasticity & then do a post-mortem examination of the residual squared $\hat{u}_i^2$ to see if they exhibit any systematic pattern. Although $\hat{u}_i^2$ are not the same thing as u_i^2, they can be

used as proxies especially if the sample size is sufficiently large.

An examination of the $\hat{u}_i^2$ may reveal the following patterns (Figure 3.2).

Here we are plotting $\hat{u}_i^2$ against the estimated Y values, $\hat{Y}_i$. Then we are finding out whether the $\hat{Y}_i$ is systematically related to $\hat{u}_i^2$. If they show some patterns, it means that there is hetroscedasticity.

Figre 3.2 Detection of Hetroscedasticity



In figure 'a', we see that there is no systematic pattern between the two variables, suggesting no hetroscedasticity is present in data. But from figures 'b' to 'e' they show some patterns and therefore there is hetroscedasticity in these data.

Formal Methods

1. **Park Test**: - Park formalized the graphical method by suggesting that σi^2 is some function of the explanatory variable Xi.

His suggested functions are

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi}$$
$$or$$
$$\ln\sigma_i^2 = \ln\sigma^2 + \beta \ln X_i + V_i$$

Since σi^2 is generally not known, Park suggested using $\hat{u}_i$, as a proxy and running following regression.

$$\ln\hat{u}_i^2 = \ln\sigma^2 + \beta \ln X_i + V_i$$
$$= \alpha + \beta \ln X_i + V_i$$

If β turn out to be statistically significant, it would suggest that Heteroscedasticity is present in the data.

Then, Park test is a two stage procedure

a) We run the OLS regression disregarding the heteroscedasticity question.

b) Run the regression

2. **Glejser Test**: - Glejser test is similar to Park test in its spirit. After obtaining the residuals $\hat{u}_i$ from the OLS regression, Glejser suggests regressing the absolute values of $\hat{u}_i$ on the X variable that is thought to be closely related with σi^2 . Glejser suggested the following functional form for this.

$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + V_i$$

In empirical and practical matters one can use Glejser approach.

3. Spearman's Rank Correlation Test:- We know that the Spearman's rank Correlation Coefficient is,

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

Where, di = difference in the rank

n= no. of individual

Assume, $Y_i = \beta_0 + \beta_1 X_i + u_i$

Then the rank correlation coefficient can be used to detect hetroscedasticity as follows:

Step 1 :- fit the regression line to the data on Y and X and obtain the residuals $\hat{u}_i$

Step 2:- taking their absolute value, $/\hat{u}_i/$, rank both $/\hat{u}_i/$ and X_i or $\hat{Y}_i$ according to an ascending or descending order and compute Spearman's rank correlation coefficient.

Step 3:- assuming that the population rank correlation coefficient Ps = 0, and n>8, the significance of the sample rank correlation r_s can be tested by the 't' test as follows,

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$ with n-2 degrees of freedom

If the computed 't' value> critical 't' value, we accept the hypothesis of hetroscedasticity. Otherwise if the computed 't' value < critical 't' value, we reject the hypothesis of hetroscedasticity assumption.

If the regression model involves more than one X variable, r_s

can be computed between /$\hat{u}_i$/ and each X variables separately and can be tested for the statistical significance using 't' test.

4. GoldFeld Quandt Test: - One of the popular methods, in which of one assumes that the Heteroscedastic variance σi^2 is positively related to one of the explanatory valuables in the regression model.

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

Suppose $\sigma i^2$is positively related to Xi as;

$$\sigma i^2 = \sigma^2 X i^2$$

Where σ^2 is a constant. This equation gave us the idea that, σi^2 is proportional to the square of the X variable. That is, σi^2 would be larger if X variable become larger. Therefore, hetroscedasticity is more likey to be present in the model.

Remedial Measures for Hetroscedasticity

- *When σi^2 is known:* ***The Method of Weighted Least Squares***

As we have seen, if σi^2 is known, the most straight forward method of correcting heteroscedasticity is by means of weighted least squares. The estimators thus obtained are BLUE. To fit this idea, consider a two variable regression model,

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{.....(1)}$$

Assume that, the true error variance σi^2 is known. That is, the error variance for each observation is known. Now consider the following transformation of the model,

$$\frac{Y_i}{\sigma_i} = \beta_1 \, 1/\sigma_i + \beta_2 \, X_i/\sigma_i + u_i/\sigma_i \text{..... (2)}$$

That is, we deflate or devide both sides of the regression model by the known σ_i. Now let, $v_i = {u_i}/{\sigma_i}$ where, v_i= the transformed error term. If v_i is homoscedastic, then the transformed regression does not suffer from the problem of hetroscedasticity. Thus it can be estimated using the usual OLS method. Assuming all other assumptions of the CLRM are fulfilled, OLS estimators of the parameters in the equation will be BLUE and we can then proceed to statistical inference in the usual manner. WLS is simply the OLS applied to the transformed model.

- *When σi^2 is not known:*

If true σi^2 are known, we can use the WLS method to obtain BLUE estimators. Since the true σi^2 are rarely known. Therefore, if we want to use the method of WLS, we will have to resort to some adhoc assumption about σi^2 and transform the original regression model so that the transformed model satisfies the hetroscedasticity assumption.

- *Re-specification of the model*

Instead of speculating σi^2, a re-specification of the model choosing a different functional form can reduce hetroscedasticity. For example, instead of running linear regression, if we estimate the model in the log form, it often reduces hetroscedasticity.

3.2 Autocorrelation

There are generally three types of data that are available for empirical analysis:

(1) Cross section,

(2) Time series, and

(3) Combination of cross section and time series, also known as pooled data.

In developing the classical linear regression model (CLRM) we made several assumptions. However, we noted that *not* all these assumptions would hold in every type of data. As a matter of fact, we saw in the previous section that the assumption of homoscedasticity, or equal error variance, may not be always tenable in cross-sectional data. In other words, cross-sectional data are often plagued by the problem of heteroscedasticity.

However, in cross-section studies, data are often collected on the basis of a random sample of cross-sectional units, such as households (in a consumption function analysis) or firms (in an investment study analysis) so that there is no prior reason to believe that the error term pertaining to sample is correlated with the error term of another sample. If by chance such a correlation is observed in cross-sectional units, it is called **spatial autocorrelation,** that is, correlation in space rather than over time.

However, it is important to remember that, in cross-sectional analysis, the ordering of the data must have some logic, or economic interest, to make sense of any determination of whether (spatial) autocorrelation is present or not. The situation, however, is likely to be very different if we are dealing with time series data, for the observations in such data follow a natural ordering over time so that successive observations are likely to exhibit inter-correlations, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year. If you observe stock price indexes it is not unusual to find that these

indexes move up or down for several days in succession. Obviously, in situations like this, the assumption of **no auto,** or **serial, correlation** in the error terms that underlies the CLRM will be violated. This situation is termed as the autocorrelation.

Here we are interested to explain,

- The nature of autocorrelation
- The reasons for autocorrelation
- Theoretical and practical consequences of autocorrelation
- The measures to detect the problem of autocorrelation and
- The measures to solve autocorrelation

3.2.1 Nature of Autocorrelation

If there are no correlation between members of series of observation ordered in time (as in time series data) or space as in cross-sectional data) is known as the assumption of no autocorrelation. That is,

Autocorrelation doesn't exist in the disturbance u_i if,

$$E(u_i, u_j) = 0, \text{ if } i \neq j$$

Otherwise, if the disturbance terms of a dataset that are ordered in time or space are correlated each other, the situation is generally termed as autocorrelation. That is,

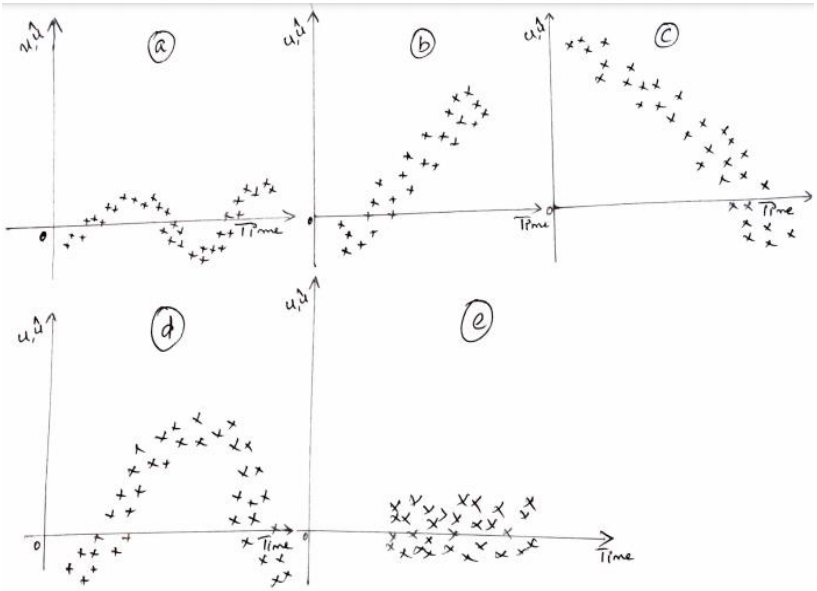
$$E(u_i, u_j) \neq 0, \text{ if } i \neq j$$

Now let us see some possible patterns of auto and no autocorrelation as Figure 3.3.

On the vertical axis of the Figure 3.3, we take both population disturbances (u) and its sample counterpart (û) and on the horizontal axis time. Then we plot the corresponding points.

From the Figure 3.3, Part (a) to Part (d) errors follow some systematic patterns. Hence, there is autocorrelation. Bute Part (e) reveals no such patterns and hence there is no autocorrelation.

Figure 3.3 Patterns of Autocorrelation

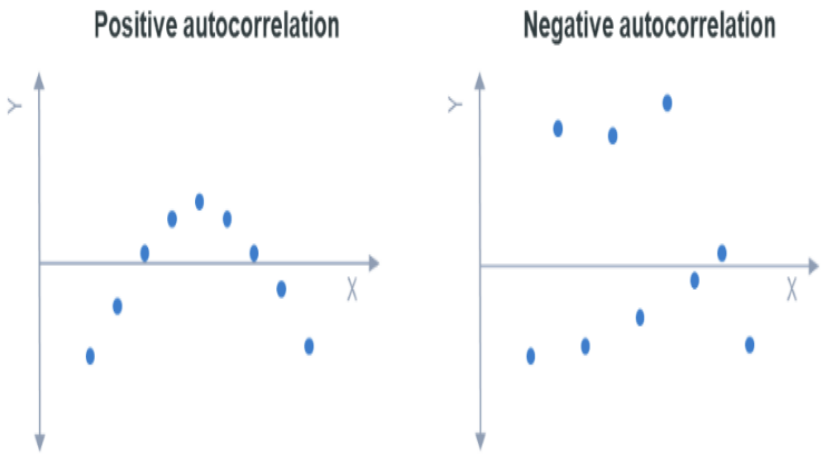


Positive and negative autocorrelation

Autocorrelation can be positive or negative. The value of autocorrelation varies from -1 (for perfectly negative autocorrelation) and 1 (for perfectly positive autocorrelation). The value closer to 0 is referred to as no autocorrelation.

Positive autocorrelation occurs when an error of a given sign between two values of time series lagged by k followed by an error of the same sign. When data exhibiting positive autocorrelation is plotted, the points appear in a smooth snake-like curve, as on the left in Figure 3.4.

Figure 3.4 Types of Autocorrelation



Negative autocorrelation occurs when an error of a given sign between two values of time series lagged by k followed by an error of the different sign. With negative autocorrelation, the points form a zigzag pattern if connected, as shown on the right of figure 3.3.

3.2.2 Reasons of Autocorrelation

The following are the major reasons for autocorrelation.

1. *Inertia*: - Silent feature of most of the time series is inertia or sluggishness. Well known examples in time series are GNI, price Index.

2. *Specification Bias: Excluded variable case*: - Residuals (which are estimate of u_i) may suggest that same variable that were originally candidates but were not included in the model for a variety of reasons should be included.

$$Y_i = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_i$$

Y_i = Quantity of beef demanded.

X_2 = Price of beef

X_3 = Consumer income

X_4 = Price of Pork

t = Time

After Regression,

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + V_t$$

3. *Specification Bias: Incorrect functional form*:-

For explaining this, first we are taking case of a marginal cost function,

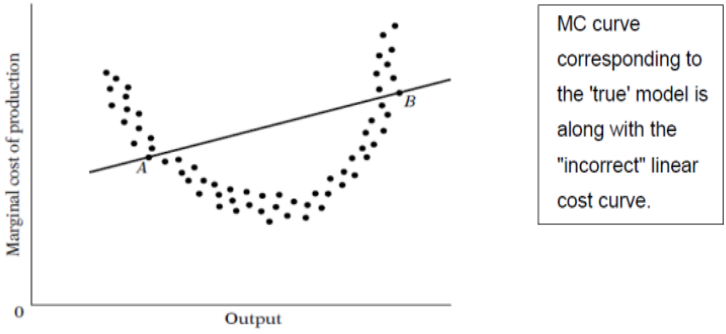
Marginal $\text{Cost}_t = \beta_1 + \beta_2\, \text{output} + \beta_3 \text{output}_i^2 + u_i$

But instead of this, suppose we get the following model.

$$MC_t = \alpha_1 + \alpha_2\, \text{output}_t + V_i$$

This can be depicted as Figure 3.5.

Figure 3.4 Specification Bias



Specification bias: incorrect functional form.

4. *Cobweb Phenomenon*: - The supply of many agricultural commodities reflects the so called cobweb Phenomenon.

Where supply reacts to price with a lag of one time period because supply decisions takes time implement.

$$Supply_t = \beta_1 + \beta_2 P_{t-1} + u_t$$

5. *Lag:* - In time series regression model, sometimes the lagged value of the dependant variable also included as one of the explanatory variable. For example,

Consumption= $\beta_1 + \beta_2$ Income + β_2 $Consumption_{t-1}$+ u_t

6. *Manipulation of data*: - In empirical analysis the raw data are often manipulated.

7. *Data Transformation*:- Sometimes data transformation leads to autocorrelation. For example,

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad \rightarrow 1$$

Y = Consumption, X = Income

$Y_{(t-1)} = \beta_1 + \beta_2 X_{(t-1)} + u_{(t-1)}$ $\rightarrow 2$ Previous Period

$Y_{(t-1)}$, $X_{(t-1)}$, $u_{(t-1)}$ are lagged values of X_1 Y & U

Sub. (II) from (I) we get

$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t$ $\rightarrow$ Δ first difference operator

For empirical purpose,

$$\Delta Y_t = \beta_2 \Delta X_t + V_t \qquad \rightarrow \qquad V_t = \Delta u_t = (u_t - u_{t-1})$$

3.2.3 Consequences of Autocorrelation

In the presence of autocorrelation one should not use OLS for estimation, to establish confidence intervals and to test hypothesis. We should use Generalised Least Squares (GLS)

method for these purposes. Because in the presence of autocorrelation,

1. The least square estimators are still linear and unbiased.
2. But they are not efficient compared to the procedures that take into account autocorrelation. In short, the usual OLS estimators are not BLUE because they do not possess the property of minimum variance.

Apart from this, the other consequences of autocorrelation are;

3. The estimated variances of OLS estimators are biased. Sometimes, the usual formulas to compute the variances and standard errors of OLS estimators seriously underestimate true variances and standard errors, there by inflating 't' values
4. Therefore, the usual 't' and F tests are not generally reliable.
5. The usual formula to compute the error variance is a biased estimator of true σ^2.
6. As a consequence, the conventionally computed R^2 may be unreliable measure of true R^2
7. The conventionally computed variances and SEs of forecast may also be inefficient.

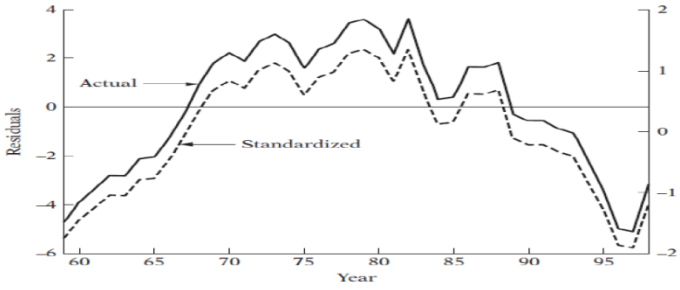
3.2.4 Detection Measures

There are varieties of tests to detect autocorrelation.

1. Graphical Method

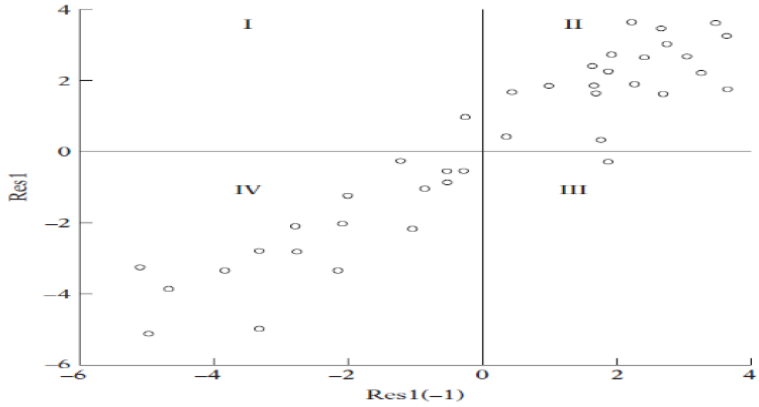
There are various ways of examine the residuals (error) under graphical method.

a) Time sequence plot (Figure 3.6)



Residuals and standardized residuals from the wages–productivity regression (

b) Standardized residual (Figure 3.7)



Current residuals versus lagged residuals.

Both figures (Figures 3.6 and 3.7) clearly shows that the residuals follow some systematic patterns and hence there is autocorrelation.

2. Runs test

Initially, we have several residuals that are negative, then there is a series of positive residuals, and then there are several residuals that are negative. If these residuals were purely random, could we observe such a pattern? Intuitively, it seems

unlikely. This intuition can be checked by the so-called runs test, sometimes also known as the Geary test, a nonparametric test. This is also a crude method.

For the Runs test, let us simply note down the signs of the residuals as * or -. Suppose we have these signs as;

(---------)(++++++++++++++++++++)(----------)

We now define a run as an uninterrupted sequence of one symbol or attribute, such as + or -. We further define the length of a run as the number of elements in it.

By examining how the runs behave in a strictly random sequence of observations, we can derive a test of randomness of runs. If there are too many runs, it means that the residuals change sign frequently, thus suggesting negative autocorrelation. Similarly, if there are too few runs, it suggests positive autocorrelation.

3. Durbin- Watson 'd' Statistic

It is one of the good methods as the *d* statistic is based on the estimated residuals, which are computed in regression analysis. It is defined as;

$$d=\frac{\sum(\hat{u}_{t-}-\hat{u}_{t-1})^2}{\sum(\hat{u}_t)^2}$$

It is simply the ratio of the sum of squared differences in successive residuals to the RSS. It is note that in the 'd' statistic, the number of observations is n -1 because one of the nation is lost in taking successive differences.

A great advantage of this statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. Because of this advantage it is now a

common practice to report the Durbin Watson 'd' along with summary statistics such as R^2, adjusted R^2, t ratio etc.

Durbin Watson 'd' statistic is based on some assumptions as

- The regression model includes an intercept term
- The explanatory variables the Xs are not stochastic or fixed in repeated sampling
- The disturbances U_t are generated by the first order autoregressive scheme
- The regression model does not include lagged values of the dependent variable as one of the explanatory variables
- There are no missing observations in the data

He expanded the formula of d statistic as follows;

$$\text{'d'} = \frac{\sum \hat{u}_t^2}{\sum \hat{u}_t^2} + \frac{\sum \hat{u}_{t-1}^2}{\sum \hat{u}_t^2} - \frac{2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

Since, $\sum \hat{u}_t^2$ and $\sum \hat{u}_{t-1}^2$ differ in only one observation they are approximately equal. Therefore setting

$\sum \hat{u}_t^2 = \sum \hat{u}_{t-1}^2$ may be written as

$$\text{'d'} \simeq 2\left(1 - \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}\right)$$

Now let us define the coefficient of autocorrelation, ρ, which can be determined with the help of the sample first-order coefficient of autocorrelation, $\hat{\rho}$

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

The d statistic become ;

$$d \simeq 2(1-\hat{\rho})$$

Since the value of ρ lies between -1 and + 1 it implies that the value of 'd' lies between 0 and 4. That is,

d will be $0 \leq d \leq 4$

because $\rho = -1 \leq \rho \leq 1$

→ $d \simeq 2 \rightarrow$ no autocorrelation

→ $d \simeq 0$ or 4 (closer) there is autocorrelation

3.2.5 Remedial Measures

1. Try to find out if the autocorrelation is pure autocorrelation or not because of the result of the mis-specification of the model.
2. Transformation of original model, so that in the transformed model we do not have the problem of (Pure) autocorrelation.
3. In case of large sample we can Newey-West method to obtain standard error of OLS estimators that are corrected for auto correlation.
4. In some situation we can continue to use the OLS method.

3.3 Multicollinearity

Another important assumption of the Classical Linear Regression Model (CLRM) is that there is no Multicollinearity among the regressors included in the multiple regression models. In practice, one rarely encounters perfect multicollinearity but cases of near or very high Multicollinearity can be found, where explanatory variables are linearly correlated in many instances.

The term multicollinearity was coined in 1934 by Ragnar Frisch in his book 'Confluence Analysis'.Because of strong interrelationships among the explanatory variables, it becomes difficult to find out how much each of these will influence the dependent variable. Usually economic variables are related in several ways and because of inter-relationship among the explanatory variables, often the statistical results gained from them are found to be ambiguous, a multicollinearity problem is said to exist. Under this section, we are explaining the nature, reasons, consequences, detection measures and ways to solve the problem of multicollinearity.

3.3.1 *Nature of Multicollinearity*

Strictly speaking, Multicollinearity refers to the existence of more than one exact linear relationships and collinearity refers to existence of a single linear relationship. But this distinction is rarely maintained in practice and multicollinearity refers to both cases. That is it meant the existence of a 'perfect' or exact linear relationship among some or all explanatory variables of a regression model. For the k variable regression involving explanatory variables X_1 X_2……. X_k (where X_1 = 1 for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following conditions are satisfied.

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots\ldots + \lambda_k X_k = 0 \quad \ldots\ldots\ldots(1)$$

Where λ_1 λ_2……λ_k are constants such that not all of them are zero simultaneously.

But the chances of obtaining a sample of values where the regressors are related in this fashion are rare in practice. Today however the term multicollinearity is used in a broader sense

to include the case of perfect multicollinearity (as equation 1) as well as the case where the X variables are inter-correlated but not perfectly so, as follows.

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots\ldots + \lambda_k X_k + V_i = 0 \quad \ldots\ldots\ldots (2)$$

Where, v_i=stochastic error term

To understand the difference between perfect and less than perfect multicollinearity in our example assume that, $\lambda_2 \neq 0$

Then equation (1) can be written as,

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} \quad \ldots\ldots\ldots\ldots (3)$$

The equation (3) shows that X2 is exactly linearly related to other variables. In this situation the coefficient of correlation between X_2 and the linear combination on the right side of the equation (3) is found to be Unity.

But in the case of less than perfect multicollinearity by assuming $\lambda_2 \neq 0$, equation (2) can also be written as,

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i \quad \ldots\ldots\ldots (4)$$

Equation (4) shows that X_2 is not an exact linear combination of other Xs because it is also determined by the stochastic error term v_i.

If multicollinearity is perfect, regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearityis less than perfect, regression coefficients although determinate, but have large standard errors (in relation to the coefficients themselves) which means

the coefficients cannot be estimated with much precision or accuracy.

3.3.2 Reasons for or Sources of Multicollinearity

The major reasons for or sources of Multicollinearity are,

1. Generally economic variables tend to move together over time. Economic magnitudes are influenced by the same factors and in consequence, once these determining factors become operative the economic variables show the same broad pattern of behaviour over time. Growth and trend factors in time series are the most serious cause of multicollinearity.
2. The use of lagged values of some explanatory variables as separate independent variables in the relationship also cause multicollinearity.
3. The data collection technique adopted, for example sampling over a limited range of values
4. Constraints on the model or in the population being sampled
5. Model specification errors
6. An over determined model, that is the model has more explanatory variables than the number of observations.

3.3.3 *Consequences of Multicollinearity*

It can be shown that even if the multicollinearity is very high the OLS estimators are still retain the property of BLUE.

Theoretical consequences

1. It is true that even in the case of high multicollinearity, the OLS estimators are unbiased, but unbiasedness is a multi sample or repeated sampling phenomenon. But this says

nothing about the properties of estimators in any given sample

2. It is true that collinearity does not destroy the property of minimum variance in the class of all linear unbiased estimators. The OLS estimators have minimum variance that is their efficient. But it does not mean that the variance of OLS estimator will necessarily be small
3. Multicollinearity is essentially a sample phenomenon in the sense that even if the X variables are not linearly related in the population they may be so related in the particular sample

For these reasons, the fact that the OLS estimators are BLUE despite multicollinearity is of little consolation in practice.

Practical consequences

1. Although BLUE, the OLS estimators have large variances and co-variances making the precision difficult.
2. Because of this, the confidence intervals tend to be much wider leading to the acceptance of these zero null hypothesis more rapidly.
3. Because of this, the‘t’ ratio of one or more coefficients tend to be statically insignificant in the case of high collinearity, the estimated standard error increase dramatically by making t values smaller. Therefore, in such cases one increasingly accept the null hypothesis
4. Although t ratios of one or more coefficients is insignificant statistically, the R^2 (the overall measure of goodness of fit) can be very high. That is, on the basis of ‘t’ test one or more of the partial slope coefficients are statistically insignificant and we accept

$H_0: \beta_2 = \beta_3 = ... \beta_k = 0$

But R^2 is so high,say 0.9, on the basis of F test one can reject H_0. That it is one of the signals of multicollinearity - insignificant 't' values, but a high overall R^2 and a significant F value.

5. The OLS estimators and their standard errors can be sensitive to small changes in the data.

3.3.4 Detection of Multicollinearity

Here we are going to detect multicollinearity. Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity but between its various degrees. Since multicollinearity refers to the degree of relationship between explanatory variables that are assumed to be non-stochastic, it is a feature of the sample and not of the population. Since multicollinearity is a sample phenomenon we do not have one unique method of detecting it for measuring its strength. But we have some rules of thumb which all the same. Some of them are;

1. High R^2 but few significant t-ratios

This is the Classic symptom of multicollinearity. If R^2 is high ($R^2 > 0.8$), the F test in most cases will reject H_0 (H_0 : β's are zero) but the individual t ratios are in significant and thus accept H_0. Although this diagnostic is sensible, its disadvantage is that it is too strong in the sense that multicollinearity is considered as the harmful only when all of the influence of the explanatory variables on Y cannot be disentangled.

2. High pair correlations among regressors

Another rule of thumb suggested is that if the pair wise or zero order correlation coefficient between two regressors is high (>0.8) then multicollinearity is a serious problem. It is also clear that high zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even the zero order correlation or simple correlations are comparatively low.

But in models involving more than two explanatory variables the simple or zero order correlation will not provide an unfailing guide in the presence of multicollinearity. In fact, if there are only two explanatory variables, the zero order correlation will suffice.

3. Examination of partial correlations

When there are more than two explanatory variables in the model we often relied on the partial correlation for detecting multicollinearity. Thus in the regression analysis of Y on X_2 , X_3 and X_4, a finding that $R^2_{1.234}$ is very high but $r^2_{12.34}$, $r^2_{13.24}$ and $r^2_{14.23}$ are comparatively low. It may suggest that the variables X_2 , X_3 and X_4 are highly inter-correlated and that at least one of these variables is excessively related with another. But there is no guarantee that the partial correlations will provide an efficient guide to multicollinearity, for it may happen that both R^2 and all the partial correlations are sufficiently high. That is, a given partial correlation may be compatible with different multicollinearity patterns.

4. Auxiliary regressions

Since multicollinearity arise because one or more of the regressors are in exact or approximately linear combinations

with other regressors. One way of finding out which X variable is related to other X variables is to regress each Xi on the remaining X variables and compute the corresponding R^2 (R^2_i). Each one of these regressions is called and auxiliary regression. Then the relationship established between F and R^2 in the variable as;

$$R^2_i = \frac{(R_{x1,x2,\ldots xk})^2 / (k-2)}{1-(R_{x1,x2,\ldots xk})^2 / (n-k+1)}$$

Follows the F distribution with (k-2) and (n-k+1) degrees of freedom.

In this equation,

n = sample size

k = number of explanatory variables

$R^2_{x1,\ x2\ldots xk}$ = the coefficient of determination in the regression of variable X_i on the remaining X variables.

If computed F value is greater than critical F_i at the chosen level of significance it means that the particular Xi is collinear with other Xs and that variable is dropped from the model

If the computed F is less than the critical Fi we say that it is not collinear with other Xs and retain that variable in the model.

But if there are several complex linear relations this curve fitting exercise may not proved to be of much value as it will be difficult to identify the separate interrelationships

Therefore one may adopt Klien's rule of thumb instead of testing all auxillary R^2 values. It is suggested that

multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2 the one that obtained from the regression of Y on Xs.

5. Eigen values and Condition Index

Eigen values and Condition Index are widely used to detect multicollinearity. From the Eigen values we can derive condition number k as,

k= Maximum Eigen Value/ Minimum Eigen Value.

The Conditional Index (CI) is,

$$\text{CI} = \sqrt{k} = \sqrt{\frac{\text{Maximum Eigen Value}}{\text{Minimum Eigen Value}}}$$

The rule of thumb for using k and CI for detecting multicollinearity is that,

If k is between 100 and 1000 there is moderate to strong multicollinearity and if k is greater than 1000 there is severe multicollinearity.

Alternatively, if Condition Index is in between 10 and 30 if there is moderate to strong multicollinearity and if Conditional Index is greater than 30 there is severe multicollinearity

6. Tolerance and Variance Inflation Factor (VIF)

For k variable regression model (Y, intercept and k–1 regressors) the variance of a partial regression coefficient is,

$$\text{Var}(\beta_j) = \left(\frac{\sigma^2}{\Sigma X j^2}\right)\left(\frac{1}{1-Rj^2}\right)$$

$$= \left(\frac{\sigma^2}{\Sigma X j^2}\right)\text{VIF}_j$$

Variance Inflation Factor (VIF) means the speed with which

the variance and co-variance increase and it can be expressed as

$$VIF = \left(\frac{1}{1-r_{23}{}^2}\right) \text{ (in three variable case)}$$

B_j = partial regression coefficient of the regressor X_j

$R_j^2 = R^2$ in the auxiliary regression of the X_j on the remaining (k-2) regressors.

R^2 increases towards unity as the collinearity of X_j with other regresses increases, the VIF also increases and in the limit it can be infinite.

Therefore, VIF can be used as an indicator of multicollinearity. The larger the value of VIF the more troublesome or collinear is the variables X_j and vice versa. As a rule of thumb, if the VIF > 10 of a variable that variable is set to be highly collinear

Tolerance can also be used to detect multicollinearity. It is defined as,

$$TOL_j = (1-R_j^2)$$

$$= 1- VIF_j$$

$TOL_j = 1$, If X_j is not correlated with other regressors

$TOL_j = 0$ If X_j is perfectly related to other regressors.

3.3.5 Remedial measures

Elimination of the effect of multicollinearity is not an easy task. There is no sure fire remedy, but there are only a few rules of thumb because it is a sample phenomenon. Besides despite near collinearity, OLS estimators still retain their BLUE property. The following are the solutions for the incidence of multicollinearity.

1. *A-Priory information*

It is possible that we can have some knowledge of the values of one or more parameters from previous empirical work. This knowledge can be profitably utilised in the current sample to reduce multicollinearity.

2. *Combining cross sectional and time series data*

Another technique to reduce the effect of multicollinearity is to combine cross-sectional and time series data, that is, pooling the data.

3. Dropping a variable(s) and specification bias

One simplest method when faced with severe multicollinearity is that to drop one of the collinear variables. Then the model becomes highly significant. But dropping a variable from the model to alleviate the problem of multicollinearity may lead to the specification bias. Hence the remedy may be worse than the disease in some situations because where as multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may seriously mislead us as to the true values of the parameters. The OLS estimators are BLUE despite near linearity.

4. *Transformation of the variables*

In Economics, we have time series data and we know that one reason for high multicollinearity between economic variables is that over time these variables tend to move in the same direction. Therefore, the transformation of the model can minimise if not solve the problem of collinearity. Commonly used transformation technique is first difference form.

5. *Additional or new data*

Multicollinearity is a sample feature not a population problem

it is possible that another sample involving the same variables collinearity may not be so serious. Sometimes simply increasing the sample size may reduce the collinearity problem. The larger the data set, the more the variations in the series that can be captured.

6. Re-thinking the model

Sometimes a model chosen for empirical analysis is not carefully thought out. Some time some important variables may be omitted or may be the functional form of the model is incorrectly chosen. However a proper specification of the model may reduce the problem of multicollinearity.

7. Other remedies

Other remedies for multicollinearity are;

- Factor analysis,
- Principal component analysis, and
- Ridge regression etc.

Module IV

Extensions of Two Variables and Dummy Variable Regression Model

This module discusses two major topics of the regression analysis namely extensions of two variable linear regression models and the dummy variable regression model. We can examine these two topics in detail here.

4.1 Extension of the two variable linear regression models

The classical linear regression model requires the parameters must be linear and the variables may or may not be linear. But we consider only models that are variable in parameters as well as in the variables. The models that are linear in parameters but not necessarily in the variables are considered under the head 'Extension of the two variable linear regression models'.

As an extension of two variable linear regressions we have mainly three models

1. The log linear model
2. Semi-log models

   -Log-lin models, and

   -Lin-log model, and
3. Reciprocal models

In all these models we are transforming non-linear models which are not linear in variables to a linear model for simplicity. Apart from these models we are familiarising

regression through origin as a special case of simple linear regression model.

4.1.1 Log linear model and measurement of elasticity

In the case of log linear model, we are transforming an exponential regression model to a linear model. For this first of all we are considering an exponential model as;

$$Y_i = \beta_1 X_i^{\beta_2} e^{u_i} \text{ .....(1)}$$

This exponential model can also be expressed in terms of logarithm as;

$$\ln (Y_i) = \ln (\beta_1) + \beta_2 \ln (X_i) + u_i \ln (e) \text{ .....(2)}$$

Here;

ln = natural logarithm whose base is 'e'

Therefore the model becomes;

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

(where, $\log_e e = 1$)

Substitute α for ln (β_1) we have,

$$\ln (Y_i) = \alpha + \beta_2 \ln (X_i) + u_i \text{ .....(3)}$$

Then the model becomes linear in parameters. The linearity can be obtained by using logarithm and hence we can apply OLS, such models are called log-log or double log or log linear models.

If the assumptions of classical linear regression model are fulfilled, the parameters of equation (3) can be obtained by the method of OLS by substituting it as;

$$Y_i^* = \alpha + \beta_2 X_i^* + u_i \text{ .....(4)}$$

The OLS estimators $\hat{\alpha}$ and $\hat{\beta}_2$ thus obtained will be BLUE

of α and β_2 respectively.

Important feature of the log linear regression model is that the slope of the coefficient of the model β_2, measures the elasticity of Y with respect to X. That is, the percentage changes in Y for a given percentage change in X. Thus, if Y represents the quantity of a commodity demanded and X its unit price, β_2 measures the price elasticity of demand which have considerable importance in economics.

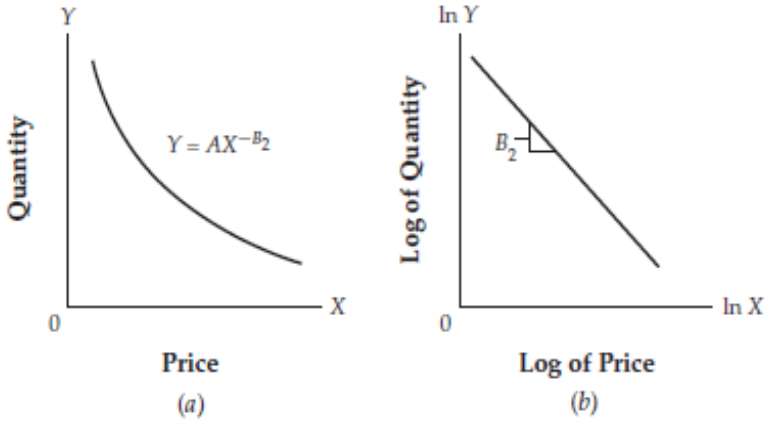
Two special features of Log linear model are;

- It is a constant elasticity model, and

- Even though $\hat{\alpha}$ and $\hat{\beta}_2$ an unbiased estimator of α and β_2 , β_1= antilog ($\hat{\alpha}$) is a biased estimator.

Constant elasticity model gives a constant total revenue change for a given percentage change in price regardless of the absolute level of price.The original model and its transformation into Log- Linear model can be differentiated as Figure 4.1.

Figure 4.1 Log-linear regression model



That is, the transformative log linear model shows constant elasticity. We can compare the original linear model and the transformed log linear model using the two slope coefficients, In linear model the slope of efficient β_2 gives the effect of a unit to change in X on the constant absolute change in Y. In log linear model the coefficient β_2 obtained from the model gives the constant percentage change in Y as a result of a 1% change in X.

We can compare the two models to compute an appropriate measure of the price elasticity. The price elasticity (E) is given as,

$$\begin{aligned} E &= \frac{\%\text{ change in } Y}{\%\text{ change in } X} \\ &= \frac{\Delta Y/Y \cdot 100}{\Delta X/X \cdot 100} \\ &= \frac{\Delta Y}{\Delta X} \cdot \frac{X}{Y} \\ &= \text{slope}\left(\frac{X}{Y}\right) \end{aligned}$$

From this,

Slope = β_2 of the linear model.

In order to obtain price elasticity we have to multiply the slope with (X/Y). But the question is that which values of X and Y are taken? If we take the average values of X and Y ($\bar{x}$ and $\bar{Y}$) for this purpose we have,

$$E = \beta_2 (\bar{x} / \bar{Y})$$

But this result is something contrasts to the price elasticity derived from the log linear model. Therefore we always depends log linear model for calculating elasticity. The basic difference between a linear model and a log linear model is that for the log linear models slope and elasticity are the same

but for a linear model,

E = Slope (X/Y)

4.1.2 Semi log models

Semi-log models included log-lin model and lin-log models

Log-lin model and measurement of growth rate

In economics, we are often interested in finding out the growth rates of GDP, population, money supply, employment, productivity, trade deficit etc. Log-lin models are very helpful in finding out these growth rates. We can explain this model as,

Suppose, we want to find out the rate of growth of real GDP over a period,

Let,

Y_t = real GDP at the time 't'

Y_0 = Initial value of real GDP

By using the compound interest formula,

$$Y_t = Y_0(1 + r)^t \quad \text{.....(1)}$$

Where, r = Compound rate of growth rate of Y

Taking natural logarithm on both sides,

$$\ln Y_t = \ln Y_0 + t \ln (1+r) \quad \text{.....(2)}$$

Substituting,

$\beta_1 = \ln Y_0$ and

$\beta_2 = \ln Y_t$

We have,

$$\ln Y_t = \beta_1 + \beta_2 t \quad \text{.....(3)}$$

Adding the disturbance term we have,

$$\ln Y_t = \beta_1 + \beta_2 t + u_t \text{.....(4)}$$

This model is also linear in parameters like any other regression models. The only difference is that the regressand is the logarithm of Y and the regressor is time which will take values 1, 2, 3 etc. That is, it is a semi log model because only one variable in the model (here Y the regressand) appears in the logarithmic form is called a log-lin model. One important property of this model is that the slope coefficient measures the constant proportional or relative change in Y for a given absolute change in the values of the regressor, time.

That is, here,

$$\beta_2 = \frac{\text{Relative change in the regressand}}{\text{Absolute change in the regressor}}$$

If we multiply the relative change in Y by 100, it gives the percentage change or growth rate in Y for an absolute change in X.

A log-lin model like equation (4) is very useful where the X variable is time in some situations such as,

β_2 = Constant relative change in the variable y

$100(\beta_2)$ = Constant percentage change in the variable Y

If $\beta_2 > 0$ = Rate of growth of variable Y

If $\beta_2 < 0$ = Rate of decay of the variable Y

That is why the models like equation (4) are called constant growth models. The growth rate obtained from log-lin models is the instantaneous rate of growth (rate of growth at a point of time). In order to calculate the compound growth rate,

Compound growth rate = $\{\text{Antilog}(\beta_2) - 1\}100$

This gives the compound growth rate over a period that we are considering for calculation.

The lin-log model

A model in which the regressand (dependent variable) is linear but the regressors are logarithmic is called a lin log model. A lin-log model can be expressed as;

$$Y_i = \beta_1 + \beta_2 \ln (X_i) + u_i \text{..... (1)}$$

Lin-Log Model is used to find the absolute change in the dependent variable for a percentage in the independent variable whereas, the log-lin model used to find the percentage growth in the dependent variable for an absolute unit change in the independent variable.

In Lin-log model,

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

We can interpret the slope coefficient β_2 as,

β_2 = Change in Y / Change in ln X

= Change in Y / Relative change in X

That is, a change in the log of a number is a relative change.

Symbolically we have;

$$B_2 = \frac{\text{absolute change in } Y}{\text{relative change in } X}$$

$$= \frac{\Delta Y}{\Delta X / X} \qquad \text{.....(2)}$$

That is,

$$\Delta Y = B_2\left(\frac{\Delta X}{X}\right) \qquad \text{.....(3)}$$

The equation (3) states that,

The absolute change in Y(ΔY) =β_2 (relative change in X)

If the later term of the equation (3) is multiplied by 100 we have the absolute change in Y for a percentage change in X.

4.1.3. Reciprocal models

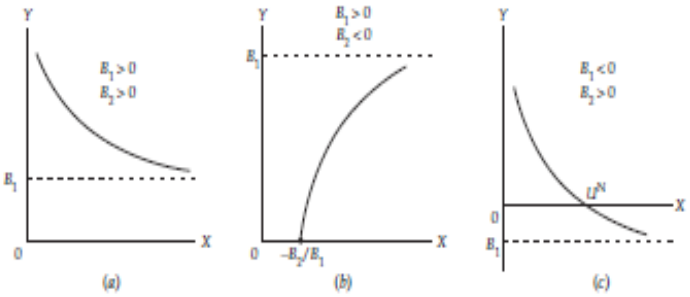
Reciprocal Models of regression model are of the following type.

$$Y_t = B_1 + B_2\left(\frac{1}{X_t}\right) + u_t \quad \text{.....(1)}$$

That is, the dependent variable Y_t is a function of the reciprocal of the independent variable X_t. This model is non-linear in variable X because it enters inversely or reciprocally, the model is linear in β_1 and β_2 and is therefore a linear regression model.

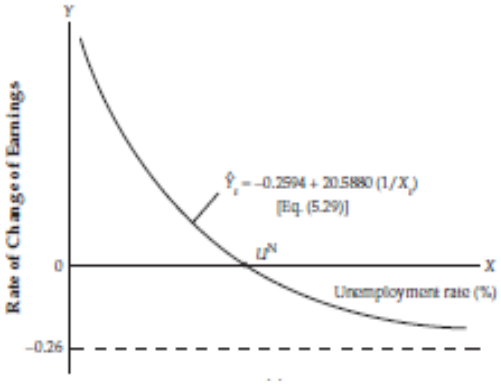
The basic feature of a reciprocal model is that as X increases indefinitely the term β_2 (1/X) approaches zero and Y approaches the limiting or asymptotic value of β_1. The reciprocal models have built in them an asymptote or limit value that the dependent variable will take when the value of the X variable increases indefinitely. Some likely shapes of the curve corresponding to the reciprocal models are shown as in Figure 4.2.

Figure 4.2 Reciprocal Models



One of the important applications of the reciprocal model is Philips curve. Using the data on percentage rate of change of money wages and employment rate for UK for the period 1861 to 1957, Phillips update a curve whose general shape is as Figure 4.3.

Figure 4.3 Phillips Curve



As the Figure 4.3 shows, there is an asymmetry in the response of wage changes to the level of unemployment rate. Wage rise faster for a unit change in unemployment rate if the unemployment rate is below U_N, which is called the natural rate of unemployment. Then fall slowly for an equivalent change when the unemployment rate is above the natural rate U_N, indicating the asymptotic floor, or -0.26, for wage change. This particular feature of the Philips curve may due to institutional factors such as union bargaining power, minimum wages, unemployment compensation etc

Since the publication of Philips' article there are different versions of Philips curve. A comparatively recent formulation is provided by Olivier Blanchard. If we let,

Π_t = Inflation rate at time t and

UN_t = Unemployment rate at time 't'

Then a modern version of Philips curve can be expressed as;

$$\Pi_t - \Pi^e_t = \beta_2 (UN_t - U^N) + U_t \text{ .....(2)}$$

Where;

Π_t =Actual inflation rate at a time t

Π^e_t =Expected inflation rate at a time 't', the expectations being formed in year t -1

UN_t =Actual unemployment rate providing at a time 't'

U^N = Natural rate of unemployment

U_t = Stochastic error term

Since Π^e_t is not directly observable, as a starting point one can make the simplifying assumption that, $\Pi^e_t = \Pi_{t-1,}$ that is, the inflation rate expected this year is the inflation rate that prevailed in the last year. Substituting this assumption in equation (2) we have ;

$$\Pi_t - \Pi_{t-1} = \beta_2 UN_t - \beta_2 U^N + U_t \text{.....(3)}$$

Writing the regression model in the standard form,

$$\Pi_t - \Pi_{t-1} = \beta_1 + \beta_2 U^N + U_t \text{.....(4)}$$

Where; $\beta_1 = \beta_2 UN_t$

The equation (4) states that the change in the inflation rate between two time periods is linearly related to the current unemployment rate. The Phillips' relation given in equation (2) is known as the 'modified Philips curve' or 'the expectation augmented Philips curve' or the 'accelerationist Philips curve'.

4.1.4 Regression through origin

There are occasions when the two variable PRF assumes the

following form;

$$Y_i = \beta_2 X_i + u_i \text{ .....(1)}$$

In this model, the intercepted term is absent or zero, hence the name regression through the origin.

For example, in Capital Asset Pricing Model (CAPM) of modern portfolio theory, the risk premium may be expressed as;

$$(ER_i - r_f) = \beta_i (ER_m - r_f) \text{ .....(2)}$$

Where;

ER_i = Expected rate of return on security 'i'

ER_m = Expected rate of return on the market portfolio

r_f = Risk free rate of return

β_i = Beta coefficient, a measure of systematic risk

If capital markets work efficiently, then capital asset pricing model postulates that security i's expected risk premium ($ER_i - r_f$) is equal to that security's β coefficient times the expected market risk premium($ER_m - r_f$).

For empirical purposes, equation (2) can be expressed as;

$$R_i - r_f = \beta_i (R_m - r_f) + u_i \text{ .....(3) or}$$

$$R_i - r_f = \alpha_i + \beta_i (R_m - r_f) + u_i \text{ .....(4)}$$

Equation (4) is known as the market model. If capital pricing model holds, α_i is expected to be zero. This form of regression is known as regression through origin.

4.2 Dummy variable regression model

In regression analysis the dependent variable is frequently influenced not only by variables that can be readily

quantifiable but also by variables that are qualitative in nature like sex, race, colour, religion, nationality etc. For example, holding all other factors constant, female workers are founded to be earning less than their male counterparts and non-whites are found to earn less than whites. This pattern may result from sex or racial discrimination, but whatever the reason qualitative variables such as sex and race do influence the dependent variable and clearly should be included among the explanatory variables.

Since qualitative variables are usually indicate the presence or absence of a quality or an attribute, such as male or female black or white, one method of quantifying such attributes is by constructing artificial variables that taken on values 1 or 0, 0 indicating the absence of an attribute and 1 indicating the presence of that attribute. For example, 1 may indicate that a person is a male and 0 may designate a female. Or 1 may indicate a person is a graduate and 0 that he is not and so on.

Variables that assume such 0 and 1 values are called dummy variables. Alternative names are'

- Indicator variables
- Binary variables
- Categorical variables
- Qualitative variables
- Dichotomies variables

4.2.1 ANOVA Models

Dummy variables can be used in regression models just as easily as quantitative variables. Here regression model contain explanatory variables that are exclusively dummy variables are called Analysis of Variance (ANOVA) models. For example,

$Y_i = \alpha + \beta D_i + u_i$.....(1)

Where;

Y_i = salary of a worker

D_i = 1, if male

= 0, if female

In (1), instead of quantitative X variable we have a dummy variable D. Model (1) enable us to find out whether sex make any difference in the salary of a worker, if all other variables such as age, education, years of experience etc. are held constant. If u_i satisfies all the assumptions of CLRM, We obtain from (1),

Mean salary of a female worker,

$$E(Y_i / D_i = 0) = \alpha \text{ .....(2)}$$

Means salary of a male worker,

$$E(Y_i / D_i = 1) = \alpha + \beta \text{ .....(3)}$$

That is, the intercepted term 'α' gives the mean salary of a female worker and the slope coefficient 'β' tells by how much the means salary of a male worker differs from the means salary of his female counterparts, $\alpha + \beta$ reflecting the mean salary of a male worker.

A test of the H_0: $\beta = 0$, that is, there is no sex discrimination can be easily made by running regression on (1) in usual manner and finding out whether on the basis of the 't' test the estimated β is statistically significant.

In most economic research, a regression model contains some explanatory variables that are quantitative and some that are qualitative. Regression models containing a mixture of quantitative and qualitative variables are called Analysis of Co-Variance (ANCOVA) models. These models can be

analyzed in details as follows.

4.2.2 ANCOVA Models

An ANCOVA model is,

$$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + u_i \text{.....(4)}$$

Where;

Y_i = salary of a worker

X_i = Years of experience

D_i = 1, if male

= 0, if female

Model (4) contains one quantitative variable (years of experience) and one qualitative variable (sex) that has two classes or categories namely male and female. (4) means that,

Mean salary of a female worker,

$$E(Y_i / X_i, D_i = 0) = \alpha + \beta X_i \text{.....(5)}$$

Means salary of a male worker,

$$E(Y_i / X_i, D_i = 1) = \alpha_1 + \alpha_2 + \beta X_i \text{.....(6)}$$

Model (4) postulates that the male and female workers salary functions in relation to the years of experience have the same slope (β) but a different intercept. In other words, it is assumed that the level of male workers' salary is different from the means salary of female workers by $\alpha_2$but the rate of change in the means salary by years of experience is same for both sexes.

If the assumption of common slope is valid, a test of the hypothesis that the two regressions (5) and (6) have the same intercept, (that is there is no sex discrimination) can be made easily by running the regression on (4) and noting the statistical significance of the estimated α_2 on the basis of 't' test. If the't' test shows that α_2 is statistically significant, we

reject the null hypothesis that the male and female workers' level of means salary are the same.

4.2.3 Dummy variable trap

To distinguish the two categories of a dummy variable, we have introduced only one dummy variable D_i. If, D_i=1, always denote a male, when D_i=0 we know that it is a female since there are only two possible outcomes. Hence one dummy variable suffices to distinguish two categories. Let us the model is as,

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \text{.....(7)}$$

Where;

Y_i = salary of a worker

X_i = Years of experience

D_{2i} = 1, if male

= 0, if female

D_{3i} = 1, if female

= 0, if male

The model (7) cannot be estimated because of perfect collinearity between D_2 and D_3. To see this, suppose we have a sample of three male workers and two female workers as follows,

	Y	α_1	D_2	D_3	X
Male	Y_1	1	1	0	X_1
Male	Y_2	1	1	0	X_2
Female	Y_3	1	0	1	X_3
Male	Y_4	1	1	0	X_4
Female	Y_5	1	0	1	X_5

It is clear from the data that,

$D_2 = 1- D_3$ or $D_3 = 1- D_2$

That is, D_2 and D_3 are perfectly collinear. In case of perfect multi-collinearity, it is clear that the usual OLS estimation is not possible. One simple way to avoid this problem is that to assign only one dummy variable if there are only two levels or classes of the qualitative variable. Thus the general rule is that if a qualitative variable has 'm' categories, introduce 'm-1' dummy variables. If this rule is not followed, we shall fall into what might be called the dummy variable trap. That is the situation of perfect multi-collinearity.

The dummy variable trap is a scenario in which the independent variables are multicolinear– a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others. In short, a dummy variable model with perfect or high multi-collinearity is the situation of 'dummy variable trap'.

4.2.4 Dummy variables and seasonal analysis

One important drawback or feature of economic time series based on monthly or quarterly data is that they exhibit seasonal patterns (regular oscillatory movement). Most of the variables were affected by seasons and it is desirable to remove the seasonal component from time series, so that one may concentrate on the trend. The process of removing the seasonal component from a time series is known as de-seasonalisation or seasonal adjustment. The time series thus obtained is called de-seasonalised or seasonally adjusted time series. Important economic time series such as the consumer price index, the wholesale price index, the index of Industrial Production are usually published in the seasonally adjusted form.

There are several methods of de-seasonalising a time series and the method of dummy variables is one of the popular methods. For this, we are using the regression equation as follows

$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} \beta X_i + u_i$.....(8)

Where;

Y_i = Profits

X_i = Sales

D_{2i} = 1, if Second quarter

= 0, If otherwise

D_{3i} = 1, Third quarter

= 0, if otherwise

D_{4i} = 1, First quarter

= 0, if otherwise

u_i = Stochastic error term

For this, firstly we have quarter-wise data and we assign values for each quarters using dummy variables. Note that we are assuming that, the variable 'season' has four classes, the four quarters of a year, thereby requiring the use of three dummy variables. Thus, if there is a seasonal pattern present in various quarters and if it is statistically significant, the estimated differential intercepts α_2, α_3, and α_4, will reflect it. It is possible that only some of these differential intercepts are statistically significant so that only some quarters may reflect it. The above model is general enough to accommodate all these cases.

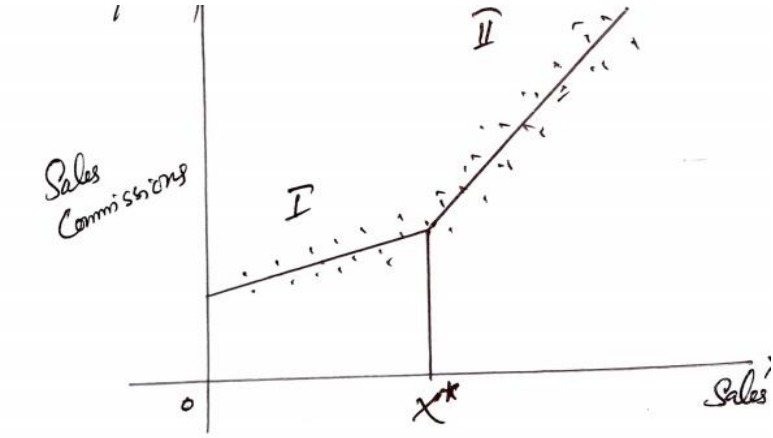
In this model it was assumed that only the intercept term differs between quarters, the slope coefficient of the sales

variable being the same in each quarter.

4.2.5 Piecewise linear regression

We can use dummy variables in another case called piecewise linear regression analysis. This case occurs when trend line occurs with different slopes. Suppose consider a case of a company remunerates its sales representatives. It pays commission based on sales in such a manner that up to a certain level the target threshold (level X*) there is one commission structure and beyond that level another. it can be depicted in the Figure 4.4.

Figure 4.4 Piece-wise regression



More specifically, it is assumed that commission for sales increases linearly with an increase in sales until the threshold level X*, after which also it increases linearly with sales but at a much steeper rate. Thus we have a piece-wise linear regression consisting of two linear pieces or segments, which are labelled I and II in the Figure 4.4 and the commission function changes its slope at the threshold value. Given the data on commission, sales and the value of the threshold level

X*, the technique of the many variables can be used to estimate the differing slopes of the two segments of the piece-wise linear regression shown in Figure4.4. Thus we proceed as,

$$Y_i = \alpha_1+\beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i \text{ ..... (9)}$$

Where;

Y_i =Sales Commission

X_i =Volume of sales

X^* = Threshold value of sales *(not known in advance)*

$D_i = 1$ if $X_i > X^*$

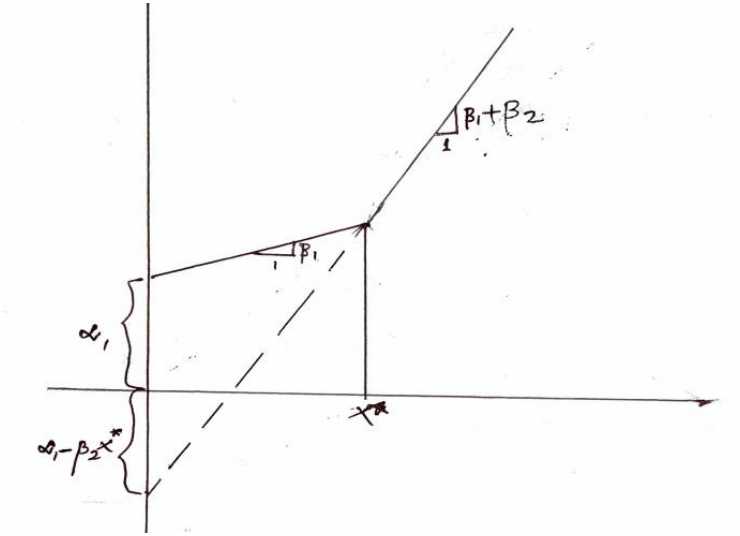
$= 0$ if $X_i < X^*$

Assuming $E(u_i) = 0$,

$E (Y_i / D_{i=1}, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1+\beta_2) X_i$ which gives the mean sales commission beyond the target level X^* and $E (Y_i / D_{i=0}, X_i, X^*) = \alpha_1 + \beta_1 X_i$ gives the mean sales commission up to the target level X^*.

Thus β_1 gives the slope of the regression line in segment I and $(\beta_1+\beta_2)$ gives the slope of the regression line segment II of the piece-wise linear regression shown in Figure 4.4. A test of the hypothesis that there is no break in the regression at the threshold value X^* can be conducted easily by noting the statistical significance of the estimator differential slope coefficient β_2 as in the Figure 4.5.

Figure 4.5 Piece-wise regression - Testing



Module V

Model Specification and Diagnostic Testing

In this module, we are discussing two important topics related to regression analysis. These are model specification errors and Qualitative Response Regression Models.

5.1 Specification Errors

One important assumption of Classical Linear Regression Model is that the regression model is correctly specified or there is no specification bias in the chosen regression model. With this assumption we are estimating the parameters of the chosen regression model and testing hypothesis about them using R^2, F, 't', etc. If the tests are satisfactory, the regression model is considered as best fit. If the tests are unsatisfactory, there are some specification errors or bias in the chosen model, such as;

- Whether some important variables are omitted from the model?
- Whether some superfluous variables included in the model?
- Is the functional form of the chosen model correct?
- Is the specification of the stochastic error correct?
- Is there more than one specification error?

If these kinds of specification errors are there, the traditional econometric methodology used is Average Economic Regression (AER).

If for example, the bias results from omission of variables, the researcher starts adding new variables to the model and tries to

'build up' the model. This traditional approach to econometric modelling is called the 'bottom-up' approach because we start our model with a given number of regressors and based on diagnostics; go on adding more variables to the model. This approach is also known as 'Average Economic Regression (AER)'

Even though so many criticisms have been raised against the Average Economic Regression, it is still have a place in the standard methodology. Here we are analysing various specification bias and how average economic regression handles the various kinds of specification errors. Before that, we are analysing how average economic regression methodology chooses a regression model first. For this, it uses the following criteria.

Parsimony:- A model can never be a completely accurate description of reality. To describe the reality, one may have to develop such a complex model that will be of little practical use. Some amount of abstractions for simplification is inevitable in any model building. The principle of parsimony states that, a model be kept as simple as possible. This means that one should introduce in the model a few key variables that capture the essence of the phenomenon under study and retain all minor and random influences to the error term u_i.

Identifiability:- For a given set of data, the identifiability means that the estimated parameters must have unique values. Or what amounts to the same thing, there is only one estimate for a given parameter.

Goodness of fit:- Since the basic thrust of regression modelling is to explain as much of the variations in the dependent variable as possible by the explanatory variables

included in the model. A model is judged good if this explanation, as measured by R^2 is high as possible.

Theoretical consistency:- A model may not be good, despite a high R^2 if one or more of the estimated coefficients have wrong signs. If for example, in the demand function if one were to obtain a positive sign for the coefficient of the price (positively sloped demand curve) one should look at that result with great suspicion. Therefore, theoretical consistency should be there when framing the models.

Predictive power:- The only relevant test of the validity of a model is comparison of its prediction with the experience. A high R^2 is used to show the predictive power of the model within the given sample. But we want is its predictive power outside the sample period

Now we are going to analyse the specification errors in detail

5.1.1 Types of specification errors

Assume that based on the theory and empirical literature, we accept a good model and let the model is,

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^{\ 2} + \beta_4 X_i^{\ 3} + u_{1i} \text{ .....(1)}$$

Where

Y = total cost of production

X = output

But suppose that for some reason a researcher decided to use the following model;

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^{\ 2} + u_{2i} \text{ .....(2)}$$

Since (1) is assumed true, adopting (2) would constitute a specification error, the error consisting in 'omitting a irrelevant variable' X_i^3. Therefore, the error term u_{2i} in (2) is in fact,

$$u_{2i} = u_{1i} + \beta_4 X_i^3 \text{ .....(3)}$$

Now suppose another researcher uses the model ,

$$Y_i = \lambda_1 + \lambda_2 X_i + \lambda_3 X_i^2 + \lambda_4 X_i^3 + \lambda_5 X_i^4 + u_{3i} \text{ .....(4)}$$

If (1) is the correct, model (4) also constitutes a specification error, the error here consisting in 'including an unnecessary or irrelevant variable' X_i^4. The new error term is in fact

$$U_{3i} = u_{1i} - \lambda_5 X_i^4 \text{ .....(5)}$$

$$= u_{1i} \text{ since, } \lambda_5 X_i^4 = 0.$$

Now assume that the model used is,

$$\ln Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 X_i^2 + \gamma_3 X_i^3 + u_{4i} \text{ .....(6)}$$

In relation to the true model (1), the model (6) would also constitute a specification bias, the bias here being the use of the wrong functional form. In (1) Y appears linearly were as in (6) it appears log linearly.

Finally consider another model,

$$Y^*_i = \beta^*_1 + \beta^*_2 X^*_i + \beta^*_3 X^{*2}_i + \beta^*_4 X^{*3}_i + u^*_{1i} \text{ .....(7)}$$

Where;

$Y^*_i = Y_i + \epsilon_i$ and

$X^*_i = X_i + w_i$ and

ϵ_i and w_i = errors of measurement

(7) states that instead of using the true Y_i and X_i we use their proxies Y^*_i and X^*_i which may contain 'errors of measurement bias'. To sum up, having once specified a model as the correct model one is likely to commit one or more of these specification errors:

1. Omission of a relevant variable

2. Inclusion of an unnecessary variable

3. Adopting the wrong functional form

4. Errors of measurement

Finally there is one more specification error which is most important it is

- Model misspecification error

This error occurs because we do not know what is the true model in the first place.

5.1.2 Consequences of specification errors

Whatever be the source of specification error, its consequences are very important. Here we are explaining two kinds of specification errors in the case of three variable regression models and this can be generalized to k-variable case.

- Omitting a relevant variable (under-fitting a model) and
- Inclusion of an irrelevant variable (over-fitting a model)

Omitting a relevant variable (under-fitting a model)

Suppose that the true model is,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \text{.....(1)}$$

But for some reasons, we fit the following model;

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i \text{.....(2)}$$

The consequences of omitting X_3 are as follows:

1. If the left out variable X_3 is correlated with the included variable X_2, $r_{23} \neq 0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are biased as well as inconsistent. That is, $E(\hat{\alpha}_1) \neq \beta_1$ and $E(\hat{\alpha}_2) \neq \beta_2$. This bias does not disappear even in large samples.

2. Even if X_2 and X_3 are uncorrelated, that is, $r_{23} = 0$, $\hat{\alpha}_1$ is still biased, although $\hat{\alpha}_2$ is unbiased

3. The var $(\hat{u}_i) = \sigma^2$ is incorrectly estimated

4. Var($\hat{\alpha}_2$) = $\sigma^2 / \Sigma X_i^2$ is a biased estimator of the variance of $\hat{\beta}_2$.

5. The usual confidence interval and hypothesis testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters

Inclusion of an irrelevant variable (over-fitting a model)

Latest assume the true model is,

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \text{.....(1)}$$

After committing the specification due to the inclusion of an unnecessary variable the model is,

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i \text{.....(2)}$$

The consequences of this specification error are as follows;

1. The OLS estimators of the parameters of the incorrect model are all unbiased and consistent. That is,

   E ($\hat{\alpha}_1$) = β_1 , E ($\hat{\alpha}_2$) = β_2 and E($\hat{\alpha}_3$) = β_3

2. The error variance σ^2 is correctly estimated
3. The usual confidence interval and hypothesis testing procedures remain valid
4. The estimated α's will be generally inefficient that is their variances will be generally larger than those of the of the true model.

5.1.3 Tests of specification errors

Once we can found that there are specification errors, there are remedies for that. Therefore, it is essential to detect whether there is any specification errors in the fitted regression model. Here we are discussing the detection measures of specification errors.

Detecting the presence of unnecessary variables

Suppose we develop a k variable model to explain a phenomenon:

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$.....(1)

Suppose we are not sure that X_k is really belongs there, the simple way to find this is to test the significance of the estimated β_k with the usual't' test $t = \hat{\beta}_k / se(\hat{\beta}_k)$. But suppose that we are not sure whether X_3 and X_4 legitimately belong in the model. In this case, we would like to test whether $\beta_3 = \beta_4 = 0$. This can be easily accomplished by the F test. Thus dictating the presence of an irrelevant variable is not a difficult task.

Tests for Omitted variables and incorrect functional form

To determine whether there is any specification bias due to omitted variables or wrong functional form the commonly used test are;

1. *Examination of residuals*

Like autocorrelation and heterosarasticity, the specification errors due to omission of a relevant variable and wrong functional form can also be detected by examining the residuals. Here also the residuals, if we plot, exhibit distinct patterns

Suppose we have a total cost function:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^{\,2} + \beta_4 X_i^{\,3} + u_{1i} \text{ .....(1)}$$

Where

Y = total cost of production

X = output

But if the researcher fits the model,

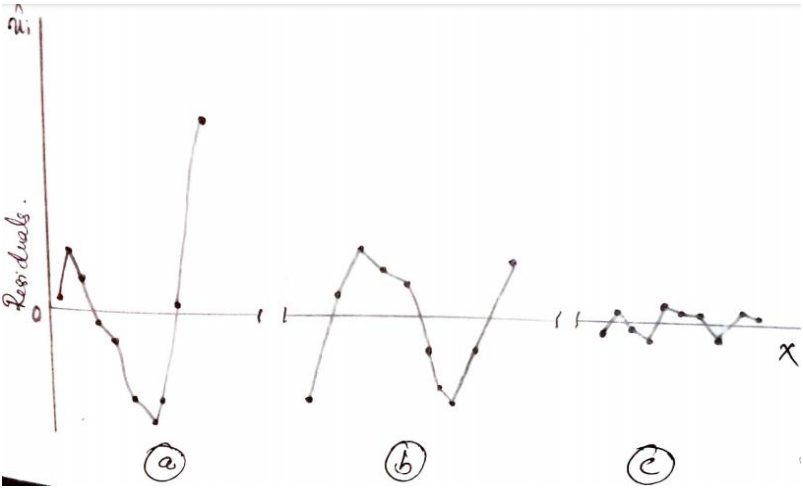
$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^{\,2} + u_{2i} \text{ .....(2)}$$

And another researcher fits the model,

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \text{ .....(3)}$$

The (2) and (3) have specification errors. If we plot the residuals we may have as Figure 5.1.

Figure 5.1 Examination of Residuals



In the Figure 5.1 as we move from left it to right ('a' to 'b' to 'c') the residuals are not only true but also they do not exhibit the pronounced cyclical swings associated with the mis-fitted models. Therefore, if there are specification errors, the residuals will exhibit noticeable patterns.

2. The Durbin Watson 'd' statistic

We are proceeding the following steps for dictating specification errors using Durbin- Watson tes.

- From the assumed model, obtain OLS residuals.
- It is assumed that this model is miss specified because it excludes at relevant explanatory variable say Z. Therefore, order the obtained residuals from step 1 according to the increasing values of Z.
- Compute the 'd' statistic from the residuals thus ordered by the usual 'd' formula ,

$$d = \frac{\sum(\hat{u}_{t-} - \hat{u}_{t-1})^2}{\sum(\hat{u}_t)^2}$$

- Base on Durbin - Watson table, if the estimated'd' value is significant, and then one can accept the hypothesis of model misspecification. Otherwise reject the hypothesis.

3. Ramsey's RESET test

Ramsey has proposed a general test of specification error called RESET (Regression Specification Error Test). Let us assume that the SLRM,

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \text{ .....(1)}$$

To detect the specification error in the model, the steps involved in RESET are ,

- From the chosen model obtain $\hat{Y}_i$ and $\hat{u}_i$.
- Plot the $\hat{u}_i$ in the graph paper to observe whether they exhibit any noticeable pattern
- If the $\hat{u}_i$ are distributed to exhibit some pattern, re-run the regression in introducing $\hat{Y}_i$ in some form as an additional regressor such as $\hat{Y}_i^2$ or $\hat{Y}_i^3$ etc. Thus we run'

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^{\,2} + \beta_4 \hat{Y}_i^{\,3} + u_i \text{ .....(2)}$$

- Obtain R^2 from (1) and (2) as R^2_{old} and R^2_{new}. Then using the F test the statistical significance of increase in R^2 using,

$$F = \frac{\frac{(R^2_{new} - R^2_{old})}{(number\ of\ new\ regressors)}}{\frac{(1-R^2_{new})}{(n-number\ of\ parameters\ in\ the\ new\ model)}}$$

- If the computed F value is significant at a 5% level one can accept the hypothesis that the model (1) is mis-specified.

5.2. Qualitative response regression models

In all the regression models that we have considered so far, we have implicitly assumed that the regressand, (the dependent variable, or the response variable) *Y* is quantitative, whereas the explanatory variables are either quantitative, qualitative (or dummy), or a mixture of both. In fact, in the previous module on dummy variables, we saw how the dummy regressors are introduced in a regression model and what role they play in specific situations. Here, we consider several models in which the regressand itself is qualitative in nature. Although increasingly used in various areas of social sciences and medical research, qualitative response regression models pose interesting estimation and interpretation challenges. In this section we are discussing some of the major themes in this area.

5.2.1 Linear probability model

To fix ideas, consider the following regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \text{ .....(1)}$$

Where,

X = family income and

Y =1 if the family owns a house and 0 if it does not own a house.

The above model looks like a typical linear regression model but because the regressant is binary, or dichotomous, it is called a Linear Probability Model (LPM). This is because the conditional expectation of Y_i given Xi, $E(Y_i/X_i)$, can be interpreted as the conditional probability that the event will occur given Xi that is, Pr ($Y_i = 1/X_i$) Thus, in our example, $E(Y_i/X_i)$ Gives the probability of a family owning a house and whose income is the given amount X_i.

The justification of the name LPM for model like (1) can be seen as follows: Assuming $E(u_i) = 0$ usual (to obtain unbiased estimators) we obtain.

$$E(Y_i/X_i) = \beta_1 + \beta_2 X_i \text{ .....(2)}$$

Now, if P_i = probability that $Y_i = 1$ (that is, the event occurs), and $(1-P_i)$ = probability that $Y_i = 0$ (that is, that the event does not occur), the variable Y_i has the following (probability) distribution.

Y_i	Probability
0	$1-P_i$
1	P_i
Total	1

That is Y_i follows the Beronoulli probability distribution.

Now, by the definition of mathematical expectation, we obtain.

$$E(Y_i) = 0(1-P_i) + 1(P_i) = P_i \text{ ----------(3)}$$

Comparing (2) with (3.), we can equate,

$$E(Y_i/X_i) = \beta_1 + \beta_2 X_i = P_i \text{ .....(4)}$$

That is, the conditional expectation of the model can, in fact, be interpreted as the conditional probability of Y_i in general; the expectation of a Bernoulli random variable is the probability that the random variable equals 1. In passing note that if there are n independent trials, each with a probability p of success and probability (1-p) of failure, and X of these trials represent the number of successes then X is said to follow the binomial distribution. The mean of the binomial distribution is 'np' and its variance is 'm(1-p)'. The term success is defined in the context of the problem.

Since the probability P_i must lie between 0 and 1, we have the restriction.

$$0 \leq E(Y_i/X_i) \leq 1 \text{----------5}$$

That is, the conditional expectation (or conditional probability) must lie between 0 and 1.

From the preceding discussion it would seem that OLS can be easily extended to binary dependent variable regression models. So, perhaps there is nothing new here. Unfortunately, this is not the case, for the LPM poses several problems, which are as follows:

Non-Normality of the Disturbances u_i.

Although OLS does not require the disturbances (u_i) to be normally distributed, we assumed them to be so distributed for the purpose of statistical inference. But the assumption of normality for u_i is not tenable for the LPMs because, like Y_i, the disturbances u_i also take only two values; that is, they also follow the Bernoulli distribution. This can be seen clearly if we write equation (1) as,

$$u_i = Y_i - \beta_1 - \beta_2 X_i \text{.....(6)}$$

The probability distribution of u_i is,

	U_i	Probability
When $Y_i=1$	$1- \beta_1- \beta_2$	Xi Pi
When $Y_i=0$	$- \beta_1- \beta_2 X_i$	1-Pi

. .....(7)

Obviously, u_i cannot be assumed to be normally distributed; they follow the Bernoulli distribution. But the non-fulfilment of the normality assumption may not be as critical as it appears because we know that the OLS point estimates still remain unbiased (recall that, if the objective is point estimation, the normality assumption is not necessary). Besides, as the sample size increases indefinitely, statistical theory shows that the OLS estimators tend to be normally distributed generally. As a result, in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption.

Heteroscedastic Variances of the Disturbances:

Even if $E(u_i) = 0$ and cov $(u_i , u_j) = 0$ for i _= j (i.e., no serial correlation), it can no longer be maintained that in the LPM the disturbances are homoscedastic. This is, however, not surprising. As statistical theory shows, for a Bernoulli distribution the theoretical mean and variance are, respectively, p and $p(1 − p)$, where p is the probability of success (i.e., something happening), showing that the variance is a function of the mean. Hence the error variance is heteroscedastic.

For the distribution of the error term given in (7), applying the definition of variance, the reader should verify that,

var $(u_i) = P_i(1 − P_i)$-------- **(8)**

That is, the variance of the error term in the LPM is heteroscedastic. Since $Pi = E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$, the variance of u_i ultimately depends on the values of X and hence is not homoscedastic.

We already know that, in the presence of heteroscedasticity, the OLS estimators, although unbiased, are not efficient; that is, they do not have minimum variance. But the problem of heteroscedasticity, like the problem of non-normality, is not insurmountable. Since the variance of u_i depends on $E(Y_i|X_i)$, one way to resolve the heteroscedasticity problem is to transform the model (1) by dividing it through by,

$$\sqrt{E(Y_i/X_i)\,[1 - E(Y_i/X_i)]} = \sqrt{P_i\,(1 - P_i)} = \text{say } \sqrt{w_i}$$

That is,

$$Y_i / \sqrt{w_i} = \sqrt{\beta_1} / w_i + \beta_2 X_i / \sqrt{w_i} + u_i / \sqrt{w_i} \text{ -------- (9)}$$

As you can readily verify, the transformed error term in (9) is homoscedastic. Therefore, after estimating (1), we can now estimate (9) by OLS, which is nothing but the *Weighted Least Squares* (WLS) with w_i serving as the weights.In theory, what we have just described is fine. But in practice the true $E(Y_i \mid X_i)$ is unknown; hence the weights w_i are unknown. To estimate w_i, we can use the following two-step procedure:

Step 1. Run the OLS regression (1) despite the heteroscedasticity problem and obtain $\hat{Y}_i$ = estimate of the true $E(Y_i \mid X_i)$. Then obtain

$$\hat{w}_i = \hat{Y}_i\,(1 - \hat{Y}_i), \text{ the estimate of } w_i.$$

Step 2. Use the estimated w_i to transform the data as shown in (9) and estimate the transformed equation by OLS (i.e., weighted least squares).

Non-fulfillment of $0 \leq E(Yi \mid X) \leq 1$

Since $E(Y_i \mid X)$ in the linear probability models measures the conditional probability of the event Y occurring given X, it must necessarily lies in between 0 and 1. Although this is true a priori, there is no guarantee that $\hat{Y}_i$, the estimators of $E(Y_i \mid X_i)$, will necessarily fulfil this restriction, and this is the real problem with the OLS estimation of the LPM. There are two ways of finding out whether the estimated $\hat{Y}_i$ lie between 0 and 1. One is to estimate the LPM by the usual OLS method and find out whether the estimated $\hat{Y}_i$ lie between 0 and 1. If some are less than 0 (that is, negative), $\hat{Y}_i$ is assumed to be zero for those cases; if they are greater than 1, they are assumed to be 1.

The second procedure is to devise an estimating technique that will guarantee that the estimated conditional probabilities $\hat{Y}_i$ will lie between 0 and 1. The logit and probit models discussed later will guarantee that the estimated probabilities will indeed lie between the logical limits 0 and 1.

5.2.2 Logit and Probit Models

As we have seen, the LPM is plagued by several problems, such as (1) non-normality of u_i, (2) heteroscedasticity of u_i, (3) possibility of $\hat{Y}_i$ lying outside the 0–1 range, and (4) the generally lower $R2$ values. But these problems are surmountable. For example, we can use WLS to resolve the heteroscedasticity problem or increase the sample size to minimize the non-normality problem. By resorting to restricted least-squares or mathematical programming techniques we can even make the estimated probabilities lie in the 0–1 interval. But even then the fundamental problem with the LPM is that it is not logically a very attractive model because it assumes that,

$P_i = E(Y = 1/X)$ increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout. This seems patently unrealistic. In reality one would expect that P_i is nonlinearly related to X_i :

At very low income a family will not own a house but at a sufficiently high level of income, say, X^*, it most likely will own a house. Any increase in income beyond X^* will have little effect on the probability of owning a house. Thus, at both ends of the income distribution, the probability of owning a house will be virtually unaffected by a small increase in X. Therefore, what we need is a (probability) model that has these two features:

(1) As *Xi* increases, $Pi = E(Y = 1 \mid X)$ increases but never steps outside the 0–1 interval, and

(2) the relationship between *Pi* and *Xi* is nonlinear, that is, "one which approaches zero at slower and slower rates as *Xi* gets small and approaches one at slower and slower rates as *Xi* gets very large.''

The reader will realize that the sigmoid, or S-shaped, curve very much resembles the **cumulative distribution function** (CDF) of a random variable. Therefore, one can easily use the CDF to model regressions where the response variable is dichotomous, taking 0–1 values. The practical question now is, which CDF? For although all CDFs are S shaped, for each random variable there is a unique CDF. For historical as well as practical reasons, the CDFs commonly chosen to represent the 0–1 response models are

(1) The logistic and

(2) The normal,

The former giving rise to the **logit** model and the latter to the **probit** (or **normit**) model.

THE LOGIT MODEL:

$$P_i = E(Y = 1/X_i) = \beta_1 + \beta_2 X_i \text{ ----------1}$$

Where X is income and Y = 1 means the family owns a house. But now consider the following representation of home ownership:

$$P_i = E(Y = 1/X_i) = \frac{1}{1+e^{-(\beta_1+\beta_2 X_i)}} \text{ ----------2}$$

For ease of exposition, we write (2) as

$$P_i = \frac{1}{1+e^{-Z_1}} = \frac{e^z}{1+e^z} \text{ ----------3}$$

Where $Z_i = \beta_1 + \beta_2 X_i$

Equation (3) represent what is known as the cumulative logistic distribution function.

It is easy to verify that as Z_i ranges from - ∞ to + ∞, P_i Ranges between 0 and 1 and that P_i is nonlinearly related to Z_i (i.e.X_i). Thus satisfying the two requirements considered earlier. But it seems that in satisfying these requirements, we have created an estimation problem because P_i is nonlinear not only in X but also in the β's as can be seen clearly from (2). This means that we cannot use the familiar OLS procedure to estimate the parameters. But this problem is more apparent than real because (2) can be linearised, which can be shown as follows.

If P_i the probability for owning a house, is given by (3) then (1-Pi), the probability of not owning a house, is

$$1 - P_i = \frac{1}{1+e^{-Z_1}} \text{ ----------4}$$

Therefore, we can write,

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i} \text{ ----------5}$$

Now $P_i / (1 - P_i)$ is simply the odds ratio in favour of owning a house, the ratio of the probability that a family will own a house to the probability that it will not own a house. Thus if $P_i = 0.8$, it means that odds are 4 to 1 in favour of the family owning a house.

Now if we take the natural log of (5) we obtain a very interesting result, namely,

$$L_i = In \frac{P_i}{1-P_i} = Z_i \text{ .....(6)}$$

$$= \beta_1 + \beta_2 X_i$$

That is, L, the log of the odds ratio, is not only linear in X, but also (from the estimation viewpoint) linear in the parameters. L is called the logit, and hence the name logit model for models like (6) Notice these features of the logit model.

1. As P goes from 0 to 1 (i.e. as Z varies from - ∞ to + ∞), the logit L goes from - ∞ to + ∞. That is, although the probabilities (of necessity) lie between 0 and 1, the logits are not so bounded.

2. Although l is linear in X, the probabilities themselves are not. This property is in contrast with the LPM model (1) where the probabilities increase linearly with X.

3. Although we have included only a single X variable, or regressor, in the preceding model, one can add as many regressors as may be dictated by the underlying theory.

4. If L, the logit, is positive, it means that when the value of the regressor(s) increases, the odds that the regressant

equals 1 (meaning some event of interest happens) increases. If L is negative, the odds that the regressant equal 1 decreases as the value of X increases. To put it differently, the logit becomes negative and increasingly large in magnitude as the odds ratio decreases from 1 to 0 and becomes increasingly large and positive as the odds ratio increases from 1 to infinity.

5. More formally, the interpretation of the logit model given in (6) is as follows: β_2, the slope, measures the change in L for unit change in X, that is, it tells how the log – odds in favor of owning a house change an income changes by a unit, say $ 1000. The intercept β_1 is the value of the log odds in favour of owning a house if income is zero. Like most interpretations of intercepts, this interpretation may not have any physical meaning.

6. Given a certain level of income, say, X, if we actually want to estimate not the odds in favor of owning a house but the probability of owning a house itself, this can be done directly from (3) once the estimate of $\beta_1 + \beta_2$ are available. This, however, raises the most important question. How do we estimate β_1 and β_2 in the first place? The answer is given in the next section.

7. Whereas the LPM assumes that Pi is linearly related to X_i the logit model assumes that the log of the odds ratio is linearly related to X_i.

THE PROBIT MODEL

The estimating model that emerges from the normal CDF is popularly normit model. To motivate the probit model, assume that in our home ownership example the decision of the i^{th} Family to own a house or not depends on an unobservable utility index I_i (also known as a latent variable), that is

determined by one or more explanatory variables, say income X_i in such a way that the larger the value of the index I_i The greater the probability of a family owning a house. We express the index I_i as.

$$I_i = \beta_1 + \beta_2 X_i \text{ ----------1}$$

Where; X_i is the income of the i^{th} Family.

How is the (unobservable) index related to the actual decision to own a house? As before let Y = 1if the family owns a house and Y = 0 if it does not. Now it is reasonable to assume that there is a critical or threshold level of the index, call it I^*_i, such that if I_i exceeds I^*_i, the family will own a house, otherwise it will not. The threshold I^*_i like I_i is not observable, but if we assume that it is normally distributed with the same mean and variance, it is possible not only to estimate the parameters of the index given in (1). But also to get some information about the unobservable index itself. This calculation is also follows.

Given the assumption of normality, the probability that I^*_i is less than or equal to I_i can be computed from the standardized normal CDF as.

$$P_i = P(Y=1/X) = P(I_i^* \leq I_i) = P(Z_i \leq \beta_1 + \beta_2 X_i) = F(\beta_1 + \beta_2 X_i) \text{ ----------2}$$

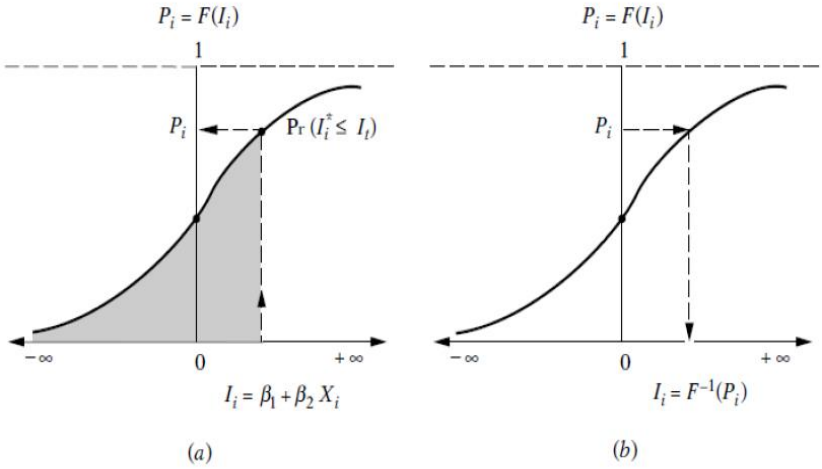
Where P(Y = 1/X) means the probability that an event occurs given the value(s) of the X, or explanatory, variable(s) and where Z*i* is the standard normal variable, i.e. $Z \sim N(0\text{-}\sigma^2)$. F is the standard normal CDF, which written explicitly in the present context is:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-z^{2/2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_i + \beta_2 X_i} e^{-z^{2/2}} dz \quad \text{.....(3)}$$

Since *P* represents the probability that an event will occur, here the probability of owning a house, it is measured by the area of the standard normal curve from -∞ as shown in Figure 5.2.

Figure 5.2 Probit model



Probit model: (*a*) given I_i, read P_i from the ordinate; (*b*) given P_i, read I_i from the abscissa.

Now to obtain information, in I_i the utility index, as well asβ_1 *and* β_2, we take the inverse of (2) to obtain.

$$I_i = F^{-1}(I_i) = F^{-1}(P_i)$$
$$= \beta_i + \beta_2 X_i \quad \text{.....(4)}$$

Where F^{-1} is the inverse of the normal CDF. What all this means can be made clear from Figure in panel a of this figure we obtain from the ordinate the (cumulative) probability of owning a house given $I^*_i \leq I_i$ whereas in panel b we obtain from the abscissa the value of I_i Given the value of P_i which is simply the reverse of the former.

In the logit model the dependent variable is the log of the odds

ratio, which is a linear function of the regressors. The probability function that underlies the logit model is the logistic distribution. If the data are available in grouped form, we can use OLS to estimate the parameters of the logit model, provided we take into account explicitly the heteroscedastic nature of the error term. If the data are available at the individual, or micro, level, nonlinear-in-the-parameter estimating procedures are called for. If we choose the normal distribution as the appropriate probability distribution, then we can use the probit model. This model is mathematically a bit difficult as it involves integrals. But for all practical purposes, both logit and probit models give similar results. In practice, the choice therefore depends on the ease of computation, which is not a serious problem with sophisticated statistical packages that are now readily available.

Address: N.H.-9, Delhi Road, Moradabad - 244001, Uttar Pradesh

Admission Helpline No. : 1800-270-1490

Contact No. :+ 91 9520 942111

Email : university@tmu.ac.in