BASIC ECONOMETRICS

Study E Material

Dr. M. Chitra

SHANLAX PUBLICATIONS

Title: **BASIC ECONOMETRICS**
Study E Material

Author's Name: **Dr. M. Chitra**

Published by: **Shanlax Publications,**
Vasantha Nagar, Madurai – 625003,
Tamil Nadu, India

Publisher's Address: **61, 66 T.P.K. Main Road,**
Vasantha Nagar, Madurai – 625003,
Tamil Nadu, India

Printer's Details: **Shanlax Press, 66 T.P.K. Main Road,**
Vasantha Nagar, Madurai – 625003,
Tamil Nadu, India

Edition Details (I,II,III): **I**

ISBN: **978-93-95422-76-5**

Month & Year: **December, 2022**

Pages: **115**

Price: ₹330

Study E - Material

BASIC ECONOMETRICS

MADURAI KAMARAJ UNIVERSITY
School of Economics
Department of Econometrics

Prepared by
Dr.M.CHITRA[1]

(For M.A Economics Programme)

[1]Dr.M.Chitra is a faculty in Department of Econometrics, studied B.Sc Mathematics, M.Sc Mathematical Economics, M.A Economics, M.Phil with Applied Econometrics, Ph.D in Economics. National and International experienced in teaching and lecturing.

Syllabus- M.A Economics
Basic Econometrics

Credits: 4 **Teaching Hours: 60**

Course Objectives

- To introduce the basic terminologies, concepts, scope and methodology of econometrics
- To equip the students with basic theory of econometrics and relevant applications of the methods

Unit I: Introduction to Basic Econometrics

Econometrics – Meaning, Features, objectives and Scope, – Tools and Types of Econometrics - Significance of stochastic disturbance term – specification of the econometric model - Methodology of Econometrics – Limitations of econometrics.

Unit II:Simple linear Regression Model

Simple linear Regression model: Specification of SLRM- OLS Estimation& Assumptions - Properties of OLS Estimators - Gauss Markov Theorem - Evaluation of SLRM : The Coefficient of determination - Application of t& f test in testing of hypothesis – interpreting and reporting the SLRM, Simple numerical problems in SLRM.

Unit III: Multiple Linear Regression Model

Elementary ideas on Multiple Linear Regression Model - Estimation - Testing of coefficients basic problems interpreting and reporting the MLRM.

Unit IV: Problems of Single Equation Models

Violation of OLS assumptions: Multicollinearity, Heteroscedasticity, Auto correlation: Meaning, sources, consequences, Detecting and remedial measures – Specification & measurement errors.

Unit V:Dynamic &Qualitative Regression Models

Instantaneous & Dynamic models- DL,AR, MA concepts – Pure and Mixed Dynamic models – Estimation of distribute lag models: Adhoc, Koyck and Almon's Approach, Regression on qualitative independent variables : ANOVA and ANCOVA models - features &advantages – regression on qualitative dependent variables: LPM&Logit models.

References

1. Baltagi, B.H. (1998), Econometrics, Springer, New York.
2. Greene, W. (1997), Econometric Analysis, New York:Prentice Hall.
3. Griffith, W.F., R.H. Hill and G.G. Judge (1993), Learning and Practicing Econometrics, New York:JohnWiley.
4. Gujrati, D. (1995), Basic Econometrics, (3rd Edition), New Delhi:McGraw Hill.
5. Intrilligator, M.D. (1978), Econometric Methods, Techniques and Applications, Prentice Hall, Englewood Cliffs, New Jersey.
6. Kmenta, J. (1997), Elements of Econometrics, Michigan Press, New York.
7. Koutsoyiannis, A. (1977), Theory of Econometrics (2nd ed.), The Macmillan Press Ltd., London.
8. Maddala, G.S. (1993), Econometrics — An Introduction, New York:McGraw-Hill.
9. Pindyck, R.S. and D.L. Rubinfeld, Econometric Models and Econometric Forecasts, 2nd Ed., McGraw-Hill Book Company, New York, 1981.
10. Wooldridge, J.M., 2013, Introductory Econometrics: A Modern Approach, NewDelhi: Cengage.

CONTENTS

UNIT I

INTRODUCTION TO BASIC ECONOMETRICS

Structure

1.1 Objectives

After reading the unit you will be able to:

- gain insights into nature of econometrics
- understand and be able to articulate, both orally and in writing, the scope of econometricsinto reality and day today life

1.2 Introduction

Econometric methods are widely used in economic research. Research required different variety of techniques, which is varied from one subject to another. In recent decades an increased emphasis has been laid down on the development and use of statistical techniques for the analysis of the economic problems. **Prof. Ragnar Frisch**, a Norwegian economist and statistician first of all named this science as "Econometrics" in 1926. Econometrics emerged as an independent discipline studying economics phenomena.

But it recognized and got attention after the world war. In 1931, the realization of the necessity of econometric work had become so evident, which made to form "Econometric Society". This International association includes practically all the worker in the field. The society published a periodical called "*Econometrica*" which disseminates the result of econometric research work. The electronic gadgets like computers have stimulated the utilization of econometrics in recent days.

1.3 Meaning and Definition

a) Meaning

Econometrics means economic measurement. Econometrics deals with the measurement of economic relationships. It's an amalgamation of economic theory with mathematics and statistics.

It is a science which combines economic theory with economic statistics and tries by mathematical and statistical methods to investigate the empirical support of general economic law established by economic theory.

The term econometrics is formed from two words of Greek origin, '*oukovouia*' meaning economy and '*uetpov*' meaning measure.

b) Definitions

The book 'Econometric Theory' was authored by**Arthur S Goldberger**, and defined econometrics in that book as"Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena".

Gerhard Tinbergen points out that "Econometrics, as a result of certain outlook on the role of economics, consists of application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results".

H Theil "Econometrics is concerned with the empirical determination of economic laws"

In the words **of Ragnar Frisch** "The mutual penetration of quantitative econometric theory and statistical observation is the essence of econometrics".

Thus, econometrics may be considered as the integration of economics, mathematics and statistics for the purpose of providing numerical values for the parameters of economic relationships and verifying economic theories. It is a special type of economic analysis and research in which the general economic theory, formulated in mathematical terms, is combined with empirical measurement of economic phenomena.

1.4 Nature of Econometrics

The nature of econometrics given in a systematic and logical way starts from its objectives, features of its equations, the uniqueness of econometrics, tools , raw materials, anatomy of econometrics, the differences of econometric models with economic models, and types of econometrics in following pages.

1.4.1 Objectives of Econometrics

The general objective of Econometrics is to give empirical content to economic theory. Empirical study means study based upon data.

- It helps to explainthe behavior of a forth coming period that is forecasting economic phenomena.
- It helps to prove the old and established relationships among the variables or between the variables
- It helps to establish new theories and new relationships.
- It helps to test the hypotheses and estimation of the parameter.

1.4.2 Features of An Equation

Econometric theory is mainly concerned with quantitative relationships among economic variables. Quantitative statements are usually expressed in the form of equation with specified numerical coefficients. Prof. Carl.F.Christ expressed that the equation must have the following features:

1. An economic equation should be **relevant** to the phenomenon being studied.
2. Equation should be **simpl**e to understand.
3. Equation should be **consistent** and consider only the relevant part of the theory.
4. Equation relating to a problem should be consistent with available **relevant** data.
5. The co-efficient of an equation will affect the economic inferences, so it is desirable to have an **accurate knowledge about the co-efficient**
6. Equation must have **forecasting ability,** because econometric study concerned with future. The above all features can be simplified as follows:

"An equation may have **relevance, simplicity, theoretical probability, explanatory ability, accuracy of co-efficient and forecasting ability**".

1.4.3 Econmetics Is a Separate Discipline. Why?

In the practice of econometrics, economic theory, institutional information and other assumptions are relied upon to formulate a statistical model, or a set of statistical hypotheses to explain the phenomena in question.

a. Economic theory makes statements or hypotheses that are mostly qualitative in nature, where Econometrics gives empirical content to most economic theory.
b. Econometrics differs from mathematical economics. The main concern of the mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. As noted above, econometrics is mainly interested in the empirical verification of economic theory. The econometrician often uses the mathematical equations proposed by mathematical economist but put these equations in such a form that they lend themselves to empirical testing.
c. Further, although econometrics presupposes the expression of economic relationships in mathematical form, like mathematical economics it does not

assume that economic relationships that are exact. On the contrary, econometrics assumes that economic relationships are not exact but stochastic. Econometric methods are designed to take into account random disturbances which create deviations from exact behavioural patterns suggested by economic theory and mathematical economics. Econometric methods are designed in such a way that they take into account the random disturbances.

d. Econometrics differs both from mathematical statistics and economic statistics. An economic statistician gathers empirical data, records them or charts them, and then attempts to describe the pattern in their development over time and detects some relationship between various economic magnitudes. Economic statistics is mainly descriptive aspect of economics. It does not provide explanations of the development of the various variables and measurement of the parameters of economic relationships.
e. On the contrary, mathematical statistics deals with methods of measurement which are developed on the basis of controlled experiments in laboratories. Statistical methods of measurement are not appropriate for economic relationships, which cannot be measured on the basis of evidence provided by controlled experiments, because such experiments cannot be designed for economic phenomena.

Econometrics uses statistical methods for adapting them to the problems of economic life. These adapted statistical methods are called econometric methods. In particular, econometric methods are adjusted so that they become appropriate for the measurement of economic relationships which are stochastic, that is, they include random elements. Hence, Econometrics is a separate discipline.

1.4.4 Tools of Econometrics

Tools of Econometrics are Mathematics and Statistics. Econometrics transforms economic theory into mathematical terms and utilizes statistical methods to derive economic relationships under certain assumptions. Algebra, properties of number system, Calculus, Statistical Data, statistical methods of sampling and testing the hypothesis are the tools of Econometrics.

1.4.5 Raw Materials of Econometrics

Data is the prime raw materials of Econometrics collected from two sources as (a) Primary and (b) Secondary. A primary source gives the first hand information and called as Primary data. The information which is already collected for some other uses is termed as Secondary data. The Secondary Data can be again classified as time series data, Cross section Data and Panel Data.

A time series is a set of observations on the values that a variable takes at different times. That is, time series data give information about the numerical values of variables from period to period. Such data may be collected at regular time intervals such as daily, weekly, monthly, quarterly, annually, quinquennially or decennially. The data thus collected may be quantitative or qualitative. Thus, data on one or more variables collected

over a period of time is called time series data. That is, values of one or more variables for several time periods pertaining to a single economic entity are given such data set is called time series data.

Cross-sectional data are data on one or more variables collected at the same point of time. These data give information on the variables concerning individual agents at a given point of time.

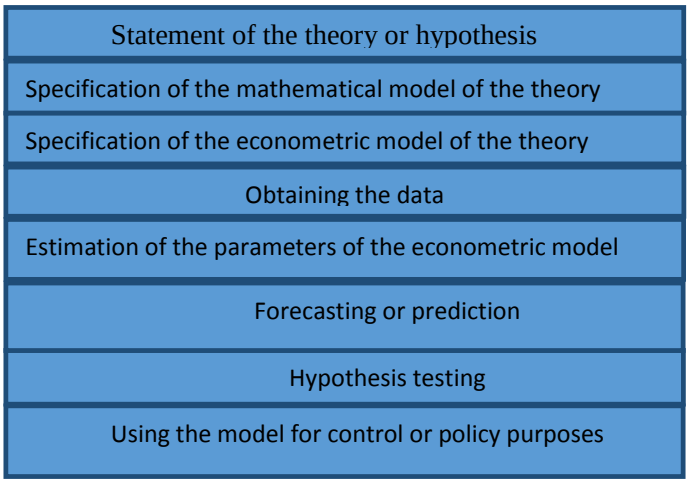
Pooled data is a combination of time series and cross sectional data. That is, in the pooled data are elements of both time series and cross-sectional data.

1.4.6 Methodolgy of Econometrics

Broadly speaking, traditional or classical econometric methodology consists of the following steps.

1) Statement of the theory or hypothesis
2) Specification of the mathematical model of the theory
3) Specification of the econometric model of the theory
4) Obtaining the data
5) Estimation of the parameters of the econometric model
6) Hypothesis testing
7) Forecasting or prediction
8) Using the model for control or policy purposes.

Flow Chart of Anatomy / Methodology of Econometrics



To illustrate the preceding steps, let us consider the well-known psychological law of consumption.

1) Statement of theory or hypothesis

Keynes stated "the fundamental psychological law.....is that men (women) are disposed, as a rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income". In short, Keynes postulated that the marginal propensity to consume (MPC), that is, the rate of change in consumption as a result of change in income, is greater than zero, but less than one. That is 0<MPC<1.

2) Specification on the mathematical model of consumption

Mathematical model is specifying mathematical equations that describe the relationships between economic variables as proposed by the economic theory. Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of functional relationship between the two. For simplicity, a mathematical economist might suggest the following form of the Keynesian consumption function:

$$Y_i = \beta_1 + \beta_2 X_i \qquad 0 < \beta_2 < 1 \qquad (1.1)$$

Where Y_i = consumption expenditure, X_i= income and $\beta_1$and β_2, known as parameters of the model are intercept and slope coefficients respectively. The slope coefficient β_2 measures the MPC.

In the above equation (1.1), the variable appearing on the left side of the equality sign is called the dependent variable and the variables on the right side are called the independent or explanatory variables. Thus, in the Keynesian consumption function, consumption expenditure is the dependent variable and income is the explanatory variable.

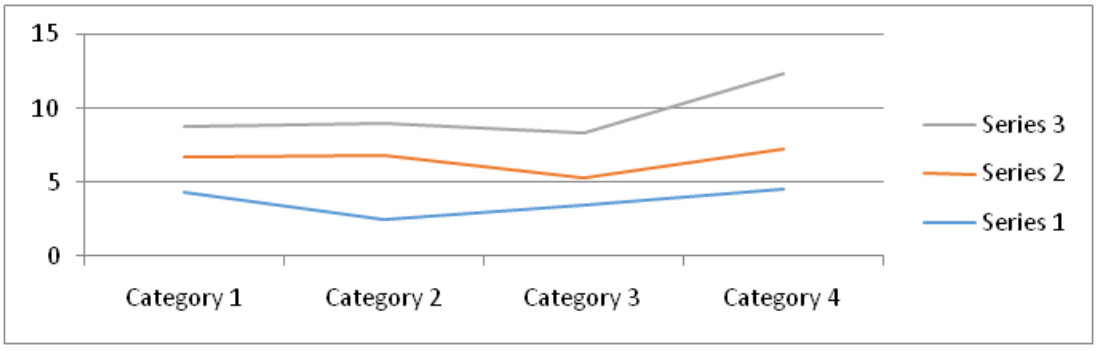
(3) Specification of the econometric model of consumption

The purely mathematical model of the consumption function as in equation (1.1) is an exact or deterministic relationship between consumption and income. But relationships between economic variables are generally inexact. This is because of the fact that in addition to income other variables affect consumption expenditure. For example, size of family, ages of the members in the family, family religion etc are likely to exert some influence on consumption.

To allow for the in exact relationship between economic variables, the econometrician would modify the deterministic consumption function as follows

$$Y_i = \beta_1 + \beta_2 X_i + U_i \qquad (1.2)$$

Where U_i is known as the disturbance or error term, which is a random or stochastic variable. The disturbance term $U_i$represents all those factors that affect consumption but are not taken into account explicitly. This equation is an example of an econometric model. More technically, it is an example of linear regression model. The econometric consumption function hypotheses that the dependent variable Y (consumption) is linearly related to the explanatory variable X (income) but that the relationship between the two is not exact; it is subject to individual variation. The econometric model of consumption function is shown as 1.2



.

(4) Obtaining Data

To estimate the econometric model given in equation (1.2), that is, to obtain the numerical values of $\beta_1$and β_2, one need data. Consumption expenditure, Income are collected from 5 respondents which are as follows

Income and expenditure of household in Madurai

S.No	Income	Consumption Expenditure
1	3000	2500
2	3500	2750
3	2750	1800
4	6500	5400
5	1780	1470
4	4000	3700
5	2570	3500

Source: Primary Data

(5) Estimation of the econometric model

After the model has been specified and data has been collected, the econometrician must proceed with its estimation. The task is to estimate the parameters of the consumption function, that is, β_1 and β_2. The numerical estimates of the parameters gives empirical content to the consumption function. Choice of the appropriate econometric technique for the estimation of the function and critical examination of the assumptions of the chosen technique is a crucial step.

(6) Hypothesis Testing

A hypothesis is a theoretical proposition that is capable of empirical verification or disproof. It may be viewed as an explanation of some event or events, and which may be true or false explanation. Confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as statistical inference or hypothesis testing.The rate of change in consumption as a result of change in income is greater than zero, but less than one. That is 0<MPC<1 will be the hypothesis.

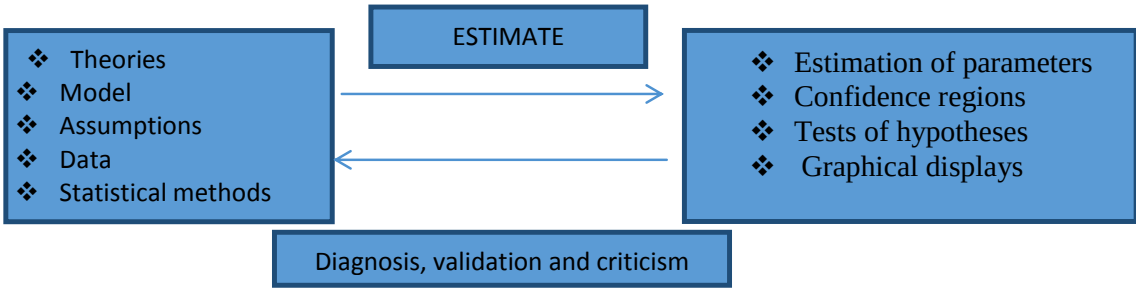
(7) Forecasting or Prediction

To predict the future values of the dependent or forecast variable Y, on the basis of known value or expected values of the explanatory, or predictor, variable X.

(8) Use the model for control or policy purposes

Suppose the estimated Keynesian consumption function, and then the government can use it for control or policy purposes such as to determine the level of income that will guarantee the target amount of consumption expenditure. In other words, an estimated model may be used for control or policy purposes. By appropriate fiscal and monetary policy mix, the government can manipulate the control variable X to produce the desired level the target variable Y.

The above process is illustrated in the following figure for better understanding:



1.4.7 Economic vs Econometric Model

Model is an abstract representation of reality which clears what is relevant to a particular question at a particular point of time and neglects all other aspects. The Economic and Econometric models study economic phenomena but two are differs in the following aspects.

(1) An Economic model is a logical representation of whatever a theoretical knowledge. A set of definition and assumption that can be used to explain a particular economic events, while
An Econometric model is an integration of endogenous variables and exogenous variables able to analyses the particular events and its spill over effects on third parties of that events.

(2) An Economic model is adapted to yield a definite and precise formulation of the economic processes at work. While
An econometric model represents a set of hypothesis that permits statistical inference from the particular data under review.

(3) Economic model needs exact and precise knowledge, while for Econometric model needs the understanding of what is relevant to the particular observations at hand.

(4) Economic model are based upon abstract economic theory; Economic theorists set great store by generality, so economic models are insufficient to permit an empirical application, While Econometric models are appropriate to the particular situation base on common sense.

(5) An economic model contains only established facts and theories while Econometric models can be introduced new relation into an economic phenomenon.

(6) Economic models are concerned with the explanation of economic laws While Econometric models are designed to forecast about economic phenomena and to serve as aids to policy formation, thee models are called as policy models.

(7) Economic models are prepared after formulation of economic laws While Construction of Econometric model is the starting point of any econometric investigation.

1.4.8 Types of Econometrics

Econometrics is divided into two broad categories as (a) Theoretical Econometrics and (b) Applied Econometrics.

Theoretical Econometrics is concerned with the development of appropriate methods for measuring economic relationship specified by econometric models. While Applied Econometrics use the tools of theoretical econometrics to study the field like economics and business specifically production function, investment function, demand and supply function etc.

1.5 Scope of Econometrics

Scope and areas of application of econometrics is expanding constantly. It includes simple as well as sophisticated mathematical and statistical techniques. Econometrics is the application of specific methods in the general field of economics science. In this sense, it plays a service role to economic analysis. By establishing new relationships and theories it serves the policy makers.

Government Aspect:

Suppose government want to devalue its currency to correct the BOP position. For estimating the consequences of devaluation, the government is concerned with price elasticity's of imports and exports. The price elasticity is to be estimated with the help of demand function of import and export commodities. Here, the econometric tools will be applied.

Producer Aspect:

Suppose a producer wants to maximize his profit, the producer will choose the level of production which gives him maximum surplus . That is minimum cost of production and maximum output, which will be solved with help of econometric methods.

In capitalistic economy too, the econometric help the producers in making rational calculations, Demand function, Price elasticity's and constraints help a producer to choose his field of investment.

Econometrics help in establish new relationships and prove old theorems. Econometrics is the outstanding method for the verification of economic theorem.

Consumer Aspect:

Effect of the taxation on consumers or effects of government expenditures on consumers standard of living are come under the purview of econometric analysis. Optimum allocation of resources has been solved with the development of the theory of programming.

Professor Oscar Lange explained the scope around three groups of questions.

(1) Earlier studies were centered round the main problem of capitalistic economy that is forecasting of business cycle. This type of study was a thing of past.

(2) Secondly econometric researches were connected with market research. Analysis of demand function, Production function, Cost function, Supply function, Distribution of wealth. etc all problems connected with market analysis.

(3) The third group of question related to theory of programming. It includes the questions relating to the whole of the economy. This field is related with planned and socialistic economies. These studies have been stimulated with the growth of communistic countries.

Now – a – days encompass mainly testing hypotheses, estimation of the parameters, usages of estimates of the parameter, ascertaining the proper functional form of economic relations, measuring the effects of imperfect data and study of the feedback relationships. Hence, whatsoever may the part of economy, or types of markets, the econometric tools are very useful for interpreting them. Whether a producer or consumer, supplier or buyer, government or public, econometrics will help in rational calculation in economic phenomena. Econometrics provides equally valuable assistance to normative as well as positive economics.

1.6 Goals of Econometrics

The three main goalsof econometrics are as follows:

1. Analysis: Econometrics primarily aims at the verification of economic theories. In this case we say that the purpose of the research is analysis. That is**, t**he economic models are formulated in an empirically testable form, to decide how well they explain the observed behavior of the economic units. Several econometric models can be derived from an economic model. Such models differ due to different choice of functional form, specification of stochastic structure of the variables etc. So, a strong analysis will be carried out by econometrics as a prime goal to verify any economic theory and economic phenomena.

2. Policy Making: The models are estimated on the basis of observed set of data and are tested for their suitability. This is the part of statistical inference of the modeling. Various estimation procedures are used to know the numerical values of the unknown parameters of the model. Based on various formulations of statistical models, a suitable and appropriate model is selected. The inference or the knowledge obtain from the numerical value of the coefficients are important for decision making of firms as well as formulation of the economic policy of the government. It helps to compare the effects of alternate policy decision.

3. Forecasting: The obtained models are used for forecasting and policy formulation which is an essential part in any policy decision. Such forecasts help the policy makers to judge the goodness of fitted model and take necessary measures in order to re-adjust the relevant economic variables.

1.7 Significance of Stochastic Disturbance Term

The disturbance term is a surrogate for all those variables that are omitted from the model but that collectively affect Y. The reasons for to introduce the stochastic disturbance termU_i are as follows:

1. **Vagueness of theory**: The theory, if any, determining the behavior of Y may be, and often is, incomplete. We might know for certain that weekly income X influences weekly consumption expenditure Y, but we might be ignorant or unsure about the other variables affecting Y. Therefore, ui may be used as a substitute for all the excluded or omitted variables from the model.

2. **Unavailability of data**: Even if we know what some of the excluded variables are and therefore consider a multiple regression rather than a simple regression, we may not have quantitative information about these variables. It is a common experience in empirical analysis that the data we would ideally like to have often are not available. For example, in principle we could introduce family wealth as an explanatory variable in addition to the income variable to explain family consumption expenditure. But unfortunately, information on family wealth generally is not available. Therefore, we may be forced to omit the wealth variable from our model despite its great theoretical relevance in explaining consumption expenditure.
3. **Core variables versus peripheral variables**: Assume in our consumption-income example that besides income X_1, the number of children per family X_2, sex X_3, religion X_4, education X_5, and geographical region X_6 also affect consumption expenditure. But it is quite possible that the joint influence of all or some of these variables may be so small and at best nonsystematic or random that as a practical matter and for cost considerations it does not pay to introduce them into the model explicitly. One hopes that their combined effect can be treated as a random variable u_i
4. **Intrinsic randomness in human behavior**: Even if we succeed in introducing all the relevant variables into the model, there is bound to be some "intrinsic" randomness in individual Y's that cannot be explained no matter how hard we try. The disturbances, the ui's, may very well reflect this 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. Consider, for example, Keynes well-known theory of the Psychological law of consumption function regards consumption expenditure (Yp) as a function of income (Xp). But since data on these variables are not directly observable, in practice we use proxy variables, such as current consumptionexpenditre (Y) and current income (X), which can be observable. Since the observed Y and X may not equal Yp and Xp, there is the problem of errors of measurement. The disturbance term u may in this case then also represent the errors of measurement. As we will see in a later chapter, if there are such errors of measurement, they can have serious implications for estimating the regression coefficients, the p's.
6. **Principle of parsimony**: Following we would like to keep our regression model as simple as possible. If we can explain the behavior of Y "substantially" with two or three explanatory variables and if our theory is not strong enough to suggest what other variables might be included, why introduce more variables? Let u_i represent all other variables. Of course, we should not exclude relevant and important variables just to keep the regression model simple.

7. **Wrong functional form**: Even if we have theoretically correct variables explaining a phenomenon and even if we can obtain data on these variables, very often we do not know the form of the functional relationship between the regressand and the regressors. Is consumption expenditure a linear (invariable) function of income or a nonlinear (invariable) function? In two-variable models the functional form of the relationship can often be judged from the scatter diagram. But in a multiple regression model, it is not easy to determine the appropriate functional form, for graphically we cannot visualize scatter diagrams in multiple dimensions.

1.7. Limitations of Econometrics

a. Applicable only to quantifiable phenomena
b. Lack of moral judgments, possibility of spurious regressions.
c. Irrational human behavior leads challenges in specifying variables and model construction for estimation.
d. Econometric model construction and data analysis are time consuming, tedious and complex because of mathematical statistical knowledge and economic theoretical knowledge are needed.
e. Data are scarce relative to the number of parameter needed to be estimated
f. Econometrics is sometime criticized for relying too heavily on the interpretation of raw data without linking it to establish economic theory or looking for casual machinist
g. It tests the hypotheses, but neglects the concerns of error

1.7 Let us Sum Up

This chapter gave the meaning, definition, aims, nature, scope, methodology of econometrics, significance of Ui term, goals of econometrics and Limitations of Econometrics.

1.8 End Chapter Exercise

A. Multiple Choice Questions

1. The book 'Econometric Theory' was authored by----------.
   A. T C Koopman B.Arther S Goldberger
   C. H. Theil D. J R N Stone
2. The name 'Econometrics' was coined by ----------.
   A. Irving Fischer B. J.M Keynes
   C. Ragnar Frisch D. Alfred Marshall
3. Econometric is an ----------.
   A. amalgamation of mathematics and statistics
   B. organization of Mathematics and Economics
   C. integration of Mathematics, Statistics and Economics
   D. integration of Economics and Statistics

4. The first step in methodology of Econometrics is ----------.
   A. Constructing Schedule C. collecting data
   B. Statement of a theory D. analyzing data.
5. The Quantitative analysis -Econometrics is based on----------.
   A. concurrent development of theory.
   B. concurrent development of observation.
   C. Both a and b.
   D. concurrent development of theory and observation, and these are related by appropriate methods of inference.
6. Why does government use econometrics?
   A. To aid decisions on prices. C. To aid decision on policy making.
   B. To aid decision on production. D. To aid profit making.
7. Why do firm owners use econometric?
   A. To aid decisions on prices.
   B. To aid decision on inventory.
   C. To aid decision on production.
   D. All of the above.
8. The term 'u' in an econometric model is usually referred to as the----------.
   A. dependent variable B. parameter
   C. error term D. hypothesis
9. The Stochastic Disturbance term Ui is introduced for representing
   A. Omitted Variable B. Vagueness of the theory
   C. Unavailability of data D. All.

B. Short and Essay Type Questions

1. What is Econometrics?What are the goals of Econometrics?
2. Differentiate the Economic model vs.Econometric model.
3. State the objectives and features of Econometrics
4. Econometrics is a separate discipline. Why?
5. What are the limitations of Econometrics?
6. Explicate the significance of stochastic error term.
7. Discuss the role of Econometrics in a Society
8. Enumerate the Methodology of Econometrics.

1.9 Reference Books

1. DamodarN.Gujarati and Sangeetha, " Basic Econometrics" Special Indian Edition , Tata McGraw Hill Education Privated Limited (Sixth Print 2010) , ISBN: 978-0-07-066005-2
2. P.G.Apte," Text book of Econometrics" Tata McGraw – Hill Publishing Company Limited
3. Dhanasekaran," Econometrics" 2nd Edition, Vrinda Publications (P) Ltd, Delhi-53 ,2011 ISBN: 978-81-8281-388-5

4. HumbertoBarreto and Frank M. Howland," Introductory Econometrics" Cambridge University Press, First South Asian Edition 2009, ISBN: 978-0-521-12358-9.
5. Dilip M. Nachane," Econometrics: Theoretical Foundations and Empirical Perspectives" Oxford University Press, Second Impression 2010 , ISBN: 978-0-19-564790-7
6. S.Shyamala,RavdeepKaur, Arul Pragasam," A text book on Econometrics: Theory and Applications" Vishal Publishing Co., Jalandhar 2017, ISBN: 81-88-646-98-9
7. S.P.Singh, Anil.K. Parashar, H.P. Singh," Econometrics and Mathematical Economics" Second Revised Editions,S.Chand and Company Ltd, New Delhi -55
8. A. Koutsoyiannis," Theory of Econometrics" Second Edition Palgrave – New York,2004,ISBN: 0-333-77822-7
9. Maddala .G.S.(1997), "Econometrics" , McGraw Hill, New York.
10. Johnston. (1997)," Econometric Methods" McGraw Hill, 4th Edition, New Delhi.

UNIT-II

SIMPLE LINEAR REGRESSION MODEL

2.1 Objectives

- To learn the procedure of estimation of Ordinary Least Square Method,
- To explore the valid interpretation of the regression estimates with assumptions on independent variables and error term
- To know the properties of estimators with proof

2.2 Introduction

We know, economics is the study of allocating the scarce resources to meet the unlimited needs and wants of human being. During the allocation process, by a mixed economy of country like India, the Demand for the goods and services or Supply of goods and services are determine by few factors directly and indirectly. If the policy makers and planners know the exact significant determinants which are influencing the demand for and supply of particular good or service, then it will be easy to allocate the resources in apt way to satisfy the consumer and supplier and study the impact into the economy of their policy implementation. Further the exact manipulated elasticity of supply and demand values will be helpful to know the scenario of an economy. Regression is tool to support in micro and macro analysis for analyzing the influencing factors, elasticity and impact of any

programme, event, policy, etc,.Hence in this chapter , an attempt is made to explain the estimation procedure, assumptions made about the independent variable and error term and properties of estimators.

2.3 Meaning of Simple Regression

Simple regression is the study of dependence of one variable (Y) with respect to another variable (X) in order to estimate and predict the unknown parameter with known values of dependent and independent variable.

2.4 The concept of Population Regression Function (PRF)

The regression Y on X is the conditional mean values of Y against the various X values. The joining of all conditional values will be resulted the regression line Y on X otherwise Population Regression Line. Hence, a Population Regression Line is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable. This can be explained by the following figure2.1 from the data table 2.1.

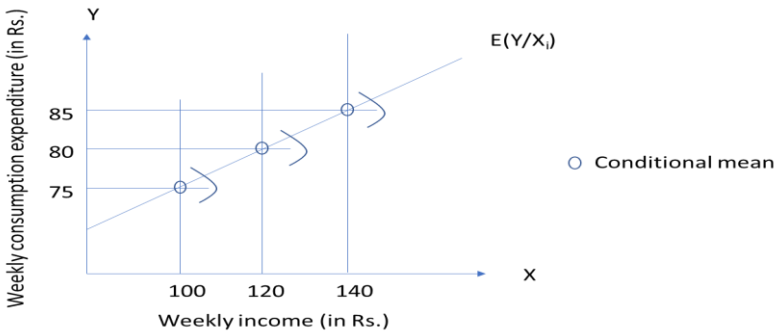


Figure 2.1 Population Regression Function

Hypothetical Example

A total population of 30 families in a village and their family income (X) and their weekly consumption expenditure (Y) given in Indian rupees. The 30 families are divided into 8 income groups from 100 to 240. Therefore 8 fixed values of X and the corresponding Y values against each of the X values in order to speak about sub population.

Table 2.1 Distribution of weekly income and weekly consumption expenditure

Y ↓	X →	100	120	140	160	180	200	220	240
Weekly family consumption expenditure Y in Rs.		65	90	85	100	120	170	200	185
		75	70	75	110	130	140	190	175
		85	80	95	115	125	150	0	145
		0	0	0	120	135	145	0	195
		0	0	0	125	140	160	0	0
Total		225	240	255	570	650	765	390	700
Conditional mean of X E(Y/Xi)		75	80	85	114	130	153	195	175

The figure 2.1 shows that for each X, there is a population of Y values, which are spread around the mean of those values. From this diagrammatic explanation, it is clear that each condition mean E(Y/Xi) is a function of X_i, where $X_i$is a given value of X

Symbolically, $E(Y/X_i) = f(x_i)$

$E(Y/X_i)$ is a linear function of X_i; it is called as ***Conditional Expectation Function***or***Population Regression Function*** (PRF). Since $E(Y/X_i)$ is a linear function of X_i, say of the type, $\boxed{E(Y/X_i) = \alpha + \beta X_i}$ where α and β are unknown parameters, where α is known as intercept and β is known as slope coefficients. The term regression, regression equation and regression model will be used synonymously.

2.5 Estimation of parameters using Ordinary Least Square (OLS) method

The method of Ordinary Least Squares is attributed to Carl Fredrisch Gauss, a German Mathematician. It is one of methods of estimation of regression analysis which is frequently used by econometricians.

The Population Regression Function is $E(Y/X_i) = \alpha + \beta X_i + u_i$, the population regression function is not directly observable. So , we have to estimate it from the sample regression function, which is as follows:The sample regression function is $\hat{Y} = \hat{\alpha} + \hat{\beta}X_i$

$Y_i = \hat{Y} + e_i$ where $\hat{Y} = \hat{\alpha} + \hat{\beta}X_i$

$\Rightarrow Y_i - \hat{Y} = e_i$(Since e_i is the difference between actual and estimated value of Y)

According to OLS assumption Σe_i^2 should be minimum.

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

$= \Sigma (Y_i - \hat{\alpha} + \hat{\beta}X_i)^2$ must be minimum ------ (1)

$$\boxed{\frac{\partial \Sigma e_i^2}{\partial(\hat{\alpha})} = \frac{\partial \Sigma e_i^2}{\partial(\hat{\beta})} = 0}$$

Equation (1) is partially derivated with respect to $\hat{\alpha}$ then

$$\therefore \frac{\partial \Sigma e_i^2}{\partial(\hat{\alpha})} = \Sigma 2(Y_i - \hat{\alpha} + \hat{\beta}X_i)(-1) = 0$$

$= -2 \Sigma (Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$

$= \Sigma Y_i - n\hat{\alpha} + \hat{\beta}\Sigma X_i = 0$

$\Rightarrow \boxed{\Sigma y_i = n\hat{\alpha} + \hat{\beta}\Sigma X_i}$ ---------- A

Equation (1) is partially derivated with respect to $\hat{\beta}$ then

$$\frac{\partial \Sigma e_i^2}{\partial(\hat{\beta})} = \Sigma 2(Y_i - \hat{\alpha} + \hat{\beta}X_i)(-x_i) = 0$$

$= -2 \Sigma (Y_i - \hat{\alpha} - \hat{\beta}X_i)x_i = 0$

$= \Sigma Y_i X_i - \hat{\alpha} \Sigma x_i - \hat{\beta} \Sigma X_i^2 = 0$

$\Rightarrow \boxed{\Sigma y_i x_i = \hat{\alpha} \Sigma x_i + \hat{\beta} \Sigma X_i^2}$ ---------- B

A and B are called as normal equation

$\Rightarrow \boxed{\Sigma y_i = n \hat{\alpha} + \hat{\beta} \Sigma X_i}$

$\Rightarrow \boxed{\Sigma y_i x_i = \hat{\alpha} \Sigma x_i + \hat{\beta} \Sigma X_i^2}$ Solving these two equation by using elimination method, one can obtain the value of $\hat{\alpha}$ and $\hat{\beta}$, The other simple alternative way is as follows:

From (A) $\Sigma Y_i = n \hat{\alpha} + \hat{\beta} \Sigma X_i$

Divide by n throughout or both sides} $\frac{\Sigma y_i}{n} = \hat{\alpha} + \hat{\beta} \frac{\Sigma X_i}{n}$

$\boxed{\overline{Y} = \hat{\alpha} + \hat{\beta} \overline{x}}$ ---------- (2) {Since $\Sigma Y_i / n = \overline{y}$ and $\Sigma X_i / n = \overline{x}$ }

From (B) normal equation, by changing its origin (0,0) to ($\overline{x}$, $\overline{y}$) we get

$\Sigma (X_i - \overline{x}) (y_i - \overline{y}) = \hat{\alpha} \Sigma (x_i - \overline{x}) + \hat{\beta} (x_i - \overline{x})^2$

Assume $\Sigma (X_i - \overline{x}) = x_i$ and $(Y_i - \overline{y}) = y_i$,then we know $\Sigma (x_i - \overline{x}) = 0$ always, therefore the first term in right hand side is equal to zero ($\hat{\alpha} \Sigma (x_i - \overline{x})$)and in the second term keeping $\hat{\beta}$ and bringing the $(x_i - \overline{x})^2$to left hand side denominator then arriving the value of $\hat{\beta}$

$\boxed{\frac{\Sigma (x_i - \overline{x}) (y_i - \overline{y})}{\Sigma (x_i - \overline{x})^2} = \hat{\beta}}$ -------- Result R_1

i.e., $\frac{\Sigma x_i y_i}{\Sigma x_i^2} = \hat{\beta}$ and substituting this $\hat{\beta}$ value in 2 we get,

$\overline{Y} = \hat{\alpha} + \frac{\Sigma x_i y_i}{\Sigma x_i^2} \overline{x}$

i.e., $\boxed{\overline{Y} - \overline{x} \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \hat{\alpha}}$ --------Result R_2

Hence, $\hat{\alpha} = \overline{Y} - \overline{x} \frac{\Sigma x_i y_i}{\Sigma x_i^2}$ and $\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$

2.6 The Classical linear Regression model: The assumptions underlying the Method of Ordinary Least Squares.

Our intention is to study the method of estimation, obtain the values of unknown parameter and draw inferences about the true parameter. In constructing the econometric

model, it is essential to depict specifically about how the independent variables and error term are created or generated for a critical valid interpretation of the regression estimators. There are ten assumptions in the context two variable regression model or Simple regression model. The assumptions are as follows:

1. The regression model is linear in the parameter. That is $Y_i = \alpha + \beta X_i + u_i$
2. X is assumed to be nonstochastic or X values are fixed in repeated sampling.
3. Given the value of X_i the mean or expected value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero.
4. Given the value of X_i,the variance of u_i is same for all observations. That is the conditional variances of u_i are identical. Technically represents the assumptions of Homoscedasticity.
5. Give any two X values, Xi and Xj (i ≠j), the correlation between any two u_i and u_j(i ≠j) is zero, where i and j are two different observations. Technically, this assumption represents that no serial correlation or autocorrelation.
6. The disturbance term u_i and explanatory variable X are uncorrelated. Technically there exist zero covariance between u_i and X_i.
7. The number of observation 'n' must be greater than the number of parameters to be estimated. Otherwise the number of observation 'n' must be greater than the number of explanatory variables.
8. The X values in a given sample must not all be the same. Technically var (X) must be a finite positive number.
9. The regression model is correctly specified. Otherwise there is no specification bias or error.
10. There is no perfect linear relationship among the explanatory variables. Technically there is no Multicollinearity.

2.7 Properties of OLS Estimators

The Ordinary least square estimators possesses the following properties

1. The estimators are linear in parameters
2. The estimators are unbiased
3. The estimators have minimum variance or least variance is known as an efficient estimator
4. The estimators which is linear, unbiased and with minimum variance or least variance is called as 'Best linear unbiased estimator' (BLUE) called Gauss Markov Theorem.
5. An unbiased estimator is said to be consistent estimator when its sample size 'n' tends to infinity and its variance tends to zero.

2.7.1 The estimators are linear in parameters

Proof: The OLS estimators are linear function of actual observation y.

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

$$= \frac{\Sigma x_i (y_i - \overline{y})}{\Sigma x_i^2}$$

$$= \frac{\Sigma x_i y_i - \Sigma x_i \overline{y}}{\Sigma x_i^2}$$

$$= \frac{\Sigma x_i y_i}{\Sigma x_i^2} - \frac{\Sigma x_i \overline{y}}{\Sigma x_i^2}$$

$$\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2} - \overline{y} \frac{\Sigma x_i}{\Sigma x_i^2}$$

$$= \frac{\Sigma x_i y_i}{\Sigma x_i^2} - 0 \qquad (\because \Sigma x_i = 0)$$

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

$\hat{\beta} = \Sigma w_i y_i$, $\hat{\beta}$ is linear. Where $\frac{\Sigma x_i}{\Sigma x_i^2} = \Sigma w_i$

Now OLS estimator of $\hat{\alpha}$ to be prove as linear, let us take

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{x}$$

$$= \frac{\Sigma y_i}{n} - \hat{\beta}\,\overline{x}$$

$$= \frac{\Sigma y_i}{n} - \Sigma w_i y_i\,\overline{x}$$

$$= \Sigma y_i \left(\frac{1}{n} - w_i\,\overline{x}\right)$$

$\hat{\alpha} = \Sigma y_i z_i$, linear ,where $z_i = \left(\frac{1}{n} - w_i\,\overline{x}\right)$

$\Sigma w_i = 0 \quad \because \Sigma x_i = 0$

$\Sigma w_i^2 = \frac{\Sigma x_i \Sigma x_i}{\Sigma x_i^2} = \frac{1}{\Sigma x_i^2}$

$\Sigma w_i x_i = \frac{\Sigma x_i \Sigma x_i}{\Sigma x_i^2} = \frac{\Sigma x_i^2}{\Sigma x_i^2} = 1$

Therefore $\hat{\alpha}$ and $\hat{\beta}$ are linear

2.7.2 Property – OLS estimators are unbiased

Proof: Let us take $\hat{\beta} = \Sigma w_i Y_i$

$= \Sigma w_i (\alpha + \beta x_i + u_i)$

$= \Sigma w_i \alpha + \Sigma w_i \beta x_i + \Sigma w_i u_i$

$= \alpha \Sigma w_i + \Sigma w_i \beta x_i + \Sigma w_i u_i$

$= \alpha . 0 + \beta \Sigma w_i x_i + \Sigma w_i u_i \qquad (\because \Sigma w_i = 0, \Sigma w_i x_i = 1)$

$= 0 + \beta + \Sigma w_i u_i$

$= \beta + \Sigma w_i u_i$

Taking expectation on both left hand side and right hand side

$E\left(\hat{\beta}\right) = E(\beta + \Sigma w_i u_i)$

$= E(\beta) + w_i E(u_i)$

$= \beta + 0 \qquad (\because E(u_i) = 0)$

$E\left(\hat{\beta}\right) = \beta$

$E\left(\hat{\beta}\right) = \beta$ is an unbiased estimator of β

Let us take, $\hat{\alpha}$ value as

$\hat{\alpha} = \overline{y} - \hat{\beta}\,\overline{x}$

Taking expectation on both sides we get

$E(\hat{\alpha}) = E(\overline{y} - \hat{\beta}\,\overline{x})$

$E(\hat{\alpha}) = E(\overline{y}) - E(\hat{\beta}\,\overline{x})$ $\quad (\because E(\overline{y}) = \alpha_o + \beta_1 \overline{x})$

$= \alpha_o + \beta_1 \overline{x} - \hat{\beta}\,\overline{x}$

$E(\hat{\alpha}) = \alpha_o$, Thus α_o is an unbiased estimator.

2.7.3. The estimators have a minimum variance or efficient estimators:

Proof: To find the variance of $\hat{\alpha}$ and $\hat{\beta}$

$\text{Var}(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta})^2]$

$= E[(\hat{\beta} - \hat{\beta})^2]$ $\quad (\because E(\hat{\beta}) = \beta)$

$= E[\Sigma w_i u_i]^2$

$= E\left[\Sigma (w_i u_i)^2 + 2 \sum_{i \neq j} \sum w_i w_j u_i u_j\right]$

$= E(\Sigma w_i u_i)^2$

$= \Sigma w_i^2 E(u_i)^2$ $\quad \left(\because E(u_i)^2 = \sigma_u^2\right) \left(\because \Sigma w_i^2 = \frac{1}{\Sigma x_i^2}\right)$

$\text{Var}(\hat{\beta}) = \sigma_u^2 \times \frac{1}{\Sigma x_i^2}$

Let us take the variance of intercept α

$\text{Var}(\hat{\alpha}) = E[(\hat{\alpha} - E(\hat{\alpha})^2]$ $\quad (\because E(\hat{\alpha}) = \alpha)$

$= E(\hat{\alpha}) - \alpha)^2$

$= E(z_i u_i)^2$

$= E\left[\Sigma z_i^2 u_i^2 + 2 \sum_{i \neq j} \sum z_i z_j u_i u_j\right]$

$= E[\Sigma z_i^2 u_i^2] + o$ $\quad (\because$ Assumption of OLS)

$= \Sigma z_i^2 E(u_i^2)$ $\quad (\because E(u_i)^2 = \sigma_u^2)$

$= \Sigma z_i^2 \sigma_u^2$

$= \sigma_u^2 \Sigma \left[\frac{1}{n} - \overline{x}\, w_i\right]^2$

$= \sigma_u^2 \Sigma \left[\frac{1}{n^2} + \bar{x}^2 w_i^2 - \frac{2}{n} \bar{x} \ w_i \right]$

$= \sigma_u^2 \left[\frac{n}{n^2} + \bar{x}^2 \Sigma w_i^2 - \frac{2}{n} \bar{x} \ \Sigma w_i \right]$ $(\because \Sigma w_i = 0)$ $\left[\Sigma w_i^2 = \frac{1}{\Sigma x_i^2} \right]$

$= \sigma_u^2 \left[\frac{1}{n} + \frac{\bar{x}}{\Sigma x_i^2} \right]$

$= \sigma_u^2 \left[\frac{\Sigma x_i^2 + n \ \bar{x}^2}{n \Sigma x_i^2} \right]$

$= \sigma_u^2 \left[\frac{\Sigma (x_i - \bar{x})^2 + n \ \bar{x}^2}{n \Sigma x_i^2} \right]$

$= \sigma_u^2 \left[\frac{\Sigma x_i^2 + \Sigma \bar{x}^2 - 2 \Sigma x_i \bar{x} + n \bar{x}^2}{n \Sigma x_i^2} \right]$ $(\because \Sigma \bar{x}^2 = n\bar{x}^2)$

$= \sigma_u^2 \left[\frac{\Sigma x_i^2 - 2n \bar{x}^2 + 2n \bar{x}^2}{n \Sigma x_i^2} \right]$ (Second term and third term in bracket are cancelled)

$= \sigma_u^2 \left[\frac{\Sigma x_i^2}{n \Sigma x_i^2} \right]$

$= \frac{\sigma_u^2}{\Sigma x_i^2} \times \frac{\Sigma x_i^2}{n}$ (x stands for multiplication of first term and second term)

$\mathbf{Var\ (\hat{\alpha}) = Var\ (\hat{\beta}) . \frac{\Sigma x_i^2}{n}}$ $\left[\because \frac{\sigma_u^2}{\Sigma x_i^2} = Var\ (\hat{\beta}) \right]$

$Var\ (\hat{\alpha}) = Var\ (\hat{\beta}) . \frac{\Sigma x_i^2}{n}$

To find the co-variance of $(\hat{\alpha})$ and $(\hat{\beta})$

$Cov\ (\hat{\alpha}\hat{\beta}) = E\ [(\hat{\alpha} - E(\alpha)(\hat{\beta} - E(\beta)]$ $[\because \hat{\alpha} = \overline{Y} - \hat{\beta}\ \overline{x}$ and $\alpha = \overline{Y} - \beta\ \overline{x}\ ]$

$= E\ [(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)]$

$Cov\ (\hat{\alpha}\hat{\beta}) = E\ [\bar{x}\ (\beta - \hat{\beta})\ (\hat{\beta} - \beta)] \Rightarrow$ Since/ $\because \hat{\alpha} - \alpha = \overline{Y} - \hat{\beta}\ \overline{x} - \overline{Y} - \beta\ \overline{x}$

$= E\ [\bar{x}\ (\beta - \hat{\beta})^2] \Rightarrow = -\bar{x}\ (\beta - \hat{\beta})$ (both Y bar cancelled

$= -\bar{x}\ E\ (\hat{\beta} - \beta)^2$

$= -\bar{x}\ Var\ \hat{\beta}$

$= -\ \bar{x}\ \frac{\sigma_u^2}{\Sigma x_i^2}$ $\left(\because Var\ \hat{\beta} = \frac{\sigma_u^2}{\Sigma x_i^2} \right)$

To prove the minimum variance of least square estimators:

Proof: Say $\hat{\beta} = \Sigma w_i Y_i$

$Var\ (\hat{\beta}) = Var\ (\Sigma w_i Y_i)$

$= \Sigma w_i Var\ Y_i$ $(\because Var\ Y_i = Varu_i = \sigma_u^2)$

$= \sigma u_i^2 \Sigma w_i^2$

$= \sigma^2 \Sigma \left(w_i - \frac{x_i}{\Sigma x_i^2} + \frac{x_i}{\Sigma x_i} \right)^2$ ⇒ The bracket terms / it's of the form $(a+b)^2$ So,

$= \sigma^2 \Sigma \left(w_i - \frac{x_i}{\Sigma x_i^2} \right)^2 + \sigma^2 \frac{\Sigma x_i^2}{(\Sigma x_i^2)^2} + 2\sigma^2 \Sigma \left(w_i - \frac{x_i}{\Sigma x_i^2} + \frac{x_i}{\Sigma x_i} \right)$

$= \sigma^2 \Sigma \left(w_i - \frac{x_i}{\Sigma x_i^2} \right)^2 + \sigma^2 \left(\frac{1}{\Sigma x_i^2} \right)$ $(\because \Sigma w_i = 0)$

$= \frac{\sigma^2}{\Sigma x_i^2}$ $\left(\because w_i = \frac{x_i}{\Sigma x_i^2} \right)$

$= \text{Var}(\hat{\beta})$

The variance of the linear estimator $\hat{\beta}$ is equal to the variance of least square estimator $\hat{\beta}$. Otherwise var $(\hat{\beta})$ >var (β). Hence, $\hat{\beta}$ is the minimum variance of linear unbiased estimator of β.

2.7.4 Gauss-Markov theorem

Statement:OLS estimators areBest Linear Unbiased Estimator (BLUE)

Proof: To show that OLS estimators are BLUE

We have shown that already in properties that

i. $\hat{\beta} = \Sigma w_i y_i$ and $E(\hat{\beta}) = \beta$ **……Linear Property**
ii. $\hat{\alpha} = \Sigma z_i y_i$ and $E(\hat{\alpha}) = \alpha$ **……Unbiased Property**
iii. $\text{Var}\,\hat{\beta} = \frac{\sigma_u^2}{\Sigma x_i^2}$ and $\text{Var}\hat{\alpha} = \sigma_u^2 \frac{\Sigma x_i^2}{n \Sigma x_i^2}$ **…..Minimum Variance Property**

Hence the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the independent variables and also unbiased estimators of $\hat{\alpha}$ and $\hat{\beta}$ respectively. Now to prove that only $\hat{\alpha}$ and $\hat{\beta}$ are also best estimators, it is essential to show that among all the unbiased estimators, the variance of the OLS estimators is the least, otherwise OLS estimators is BLUE.

Let $\hat{\hat{\beta}} = \Sigma c_i y_i$ be any other linear estimator of β, where $c_i = w_i + d_i$ and d_i being any arbitrary constant other than zero. Then to prove an unbiased estimator as follows:

$E(\hat{\hat{\beta}}) = \beta$ $(\because \hat{\hat{\beta}}$ is an unbiased estimator of $\beta)$

$E(\hat{\hat{\beta}}) = E(\Sigma c_i y_i)$

$= E(\Sigma c_i (\alpha + \beta x_i + u_i)]$

$= E(\Sigma c_i \alpha + \beta \Sigma c_i x_i + \Sigma u_i c_i)]$

$= \Sigma c_i E(\alpha) + E(\beta) \Sigma c_i x_i + \Sigma c_i E(u_i)$

$E(\hat{\hat{\beta}}) = \beta$……Unbiased Proved. (only if $\Sigma e_i=0$, and $\Sigma x_i c_i=1$, $E(u_i)=0$)

Rough work :i.e., If $\Sigma(w_i+d_i) = 0$ and $\Sigma (w_i+d_i) x_i = 1$

⇒ $\Sigma d_i = 0$, $\Sigma w_i x_i + \Sigma d_i x_i = 1$

$\Rightarrow 1+\Sigma d_i x_i=1$ $(\because \Sigma w_i x_i=1)$

$\Rightarrow \Sigma d_i x_i=1-1=0$

Now ,To prove the minimum variance property Let us take var $(\hat{\hat{\beta}}) = E[\Sigma c_i u_i]^2$

Var $(\hat{\hat{\beta}}) = E[\Sigma c_i u_i]^2$

$= E[\Sigma c_i^2 u_i^2]$

$= \sigma_u^2 \Sigma c_i^2$ $(\because E(u_i^2) = \sigma_u^2)$

$= \sigma_u^2 \Sigma (w_i+d_i)^2$

$= \sigma_u^2 (\Sigma w_i^2 + \Sigma d_i^2 + 2 \Sigma w_i d_i)$

$= \sigma_u^2 \Sigma w_i^2 + \sigma_u^2 \Sigma d_i^2 + \sigma_u^2 2 \Sigma w_i d_i$ Now $\Sigma d_i w_i = \frac{\Sigma d_i x_i}{\Sigma x_i^2}$

$= \text{Var}\,\hat{\beta} + \sigma_u^2 \Sigma d_i^2 + \sigma_u^2 2 \Sigma w_i d_i$ where $\Sigma d_i x_i = \Sigma d_i (x_i - \bar{x})$

$= \text{Var}\,\hat{\beta} + \sigma_u^2 \Sigma d_i^2 + 0$ $= \Sigma d_i x_i - \Sigma d_i \bar{x} = 0$

$= \text{Var}\,\hat{\beta} + \sigma_u^2 \Sigma d_i^2$

Hence Var $(\hat{\hat{\beta}})$ = Var $\hat{\beta}$ + a positive quantity

$\therefore$Var $(\hat{\hat{\beta}})$ - Var $\hat{\beta}$ > 0 (or) Var $(\hat{\hat{\beta}})$ >Var $\hat{\beta}$

Thus, the variance of the OLS estimators is the least among all linear unbiased estimators.

Similarly, for $\hat{\alpha}$

Let $\hat{\hat{\alpha}} = \Sigma c_i^* y_i$ be any other linear unbiased estimator of α where $c_i^* = w_i^* + d_i^*$, d_i^* being any arbitrary constant other than zero.

Then,

$\hat{\hat{\alpha}} = \Sigma c_i^* y_i$**……. Linear and to prove unbiased, Let us take**

$\hat{\hat{\alpha}} = \Sigma c_i^* (\alpha + \beta x_i + u_i)]$

$= \Sigma c_i^* \alpha + \beta \Sigma c_i^* x_i + \Sigma u_i c_i^*$ (only if $\Sigma c_i^*=1$ and $\Sigma c_i^* x_i = 0$ and $E(u_i) = 0$)

Taking expectation on both sides

$E(\hat{\hat{\alpha}}) = E(\alpha) + \Sigma c_i^* E(u_i)$ we get

$\hat{\hat{\alpha}} = \alpha + \Sigma c_i^* E(u_i)$ (since $E(u_i) = 0$)

$= \alpha + 0$

$E(\hat{\hat{\alpha}}) = \alpha$

Rough work : if $\Sigma c_i^* = 1 = \Sigma(w_i^*+d_i^*) = 1+\Sigma d_i^* \Rightarrow \Sigma d_i^* = 0$ where $\Sigma w_i^* = 1$

$\Sigma c_i^* x_i = \Sigma(w_i^*+d_i^*) x_i = \Sigma w_i^* x_i + \Sigma d_i^* x_i = 0$ $(\because \Sigma w_i^* x_i = 0, \Sigma d_i^* x_i = 0)$

$\Sigma w_i^* x_i = \Sigma d_i^* x_i + 0$

$1 = \Sigma d_i^* x_i \Rightarrow \Sigma d_i^* x_i = -1$

$$\text{Var}(\hat{\hat{\alpha}}) = E[\Sigma c_i^* u_i]^2$$
$$= E[\Sigma c_i^{*2} u_i^2]$$
$$= \sigma_u^2 \Sigma c_i^{*2}$$
$$= \sigma_u^2 \Sigma (w_i^* + d_i^*)^2$$
$$= \sigma_u^2 \Sigma w_i^{*2} + \sigma_u^2 \Sigma d_i^{*2} + \sigma_u^2 2 \Sigma w_i^* d_i^*$$
$$= \text{var}(\hat{\alpha}) + \text{a positive quantity}$$

$\text{Var}(\hat{\hat{\alpha}}) > \text{var}(\hat{\alpha})$

$\hat{\alpha}$is a linear unbiased estimator. The OLS estimators have the least variance, which is the best linear unbiased estimator (BLUE).

2.8. Goodness of Fit

Goodness of fit refers the summary measure of how well the sample regression line fits the data. Goodness of fit called as coefficient of determination (r^2) means that the proportion or percentage of the total variation in Y explained by the regression model. The properties of r^2 are (a) Non negative quantity and (b) lies between 0 to 1.If the r^2 value is equal to one then it is perfect fit. If the value of r^2 is zero means then there is relationship between the regressor and regressand or no relationship between the dependent variable and the independent variable. The coefficient of determination is the ratio of Explained Sum Square with respect to Total Sum Square. It tells about what proportion of the variation in the dependent variable or regressand is explained by the explanatory variable or regressor

2.9 Tests of Hypotheses

The theory of hypothesis testing is concerned with developing rules or procedure for deciding whether to reject or not reject the null hypothesis devised. There are two mutually complementary approached for devising such rules, namely confidence interval and test of significance. Both these approaches predict that the variable statistic or estimator.

Confidence interval approach is the concept of interval estimation. An interval estimator is an interval or range constructed in such a manner that it has a specified probability of including within it limits the true value of unknown parameter. The interval thus constructed is known as confidence interval which is often stated in percent form such as 90 percent or 95 percent.

In the significance test procedure, one develops a test statistic and examines its sampling distribution under null hypothesis. The test statistics are usually a well-defined probability distribution such as normal, t, F or chi-square. Once the test statistic is calculated or computed from the data, the its ρ value will be easily taken from the statistical tables. If the ρ value is small one can reject the null hypothesis. In choosing the ρ value the investigator has to bear in mind the probabilities of committing Type I and Type II error.

2.10 Solved Numerical Problems

Illustration: 1. Using the following data-

Investment	(Y)		65	57	57	54	66
Change in output (X)		26	13	16	-7	27	

Estimate the regression line $Y=\alpha+\beta X$, test the hypothesis that $\beta=0$ against the alternative $\beta<0$ at 5% level of significance, also construct 95% confidence interval for β.

Solution: The estimated line is $Y=\hat{\alpha}+\hat{\beta}X$. Now first of all we will calculate the parameters of the equation.

Calculation of Parameters and Error Term

X	Y	X^2	XY	$\hat{Y}$	$e=Y-\hat{Y}$	e^2
26	65	676	1690	54.55+.35x26=63.65	1.35	1.82
13	57	169	741	54.55+.35x13=59.10	-2.10	4.41
16	57	256	912	54.55+.35x16=60.15	-3.15	9.92
-7	54	49	-378	54.55+.35x-7=52.10	1.90	3.36
27	66	726	1782	54.55+.35x27=64.00	2.00	4.00
75	**299**	**1879**	**4747**	**299**		**23.51**

$$\hat{\beta}_{yx}=\frac{\sum XY-\frac{\sum X.\sum Y}{n}}{\sum X^2-\frac{\left(\sum X\right)^2}{n}}=\frac{4747-\frac{75x299}{5}}{1879-\frac{(75)^2}{5}}$$

$$=\frac{4747-4485}{1879-1125}=\frac{262}{754}=.35(approx.)$$

and also $\overline{X}$ =15 and $\overline{Y}$ =59.8

Since the equation is $Y=\hat{\alpha}+\hat{\beta}X$ on passing through mean $\overline{Y}=\hat{\alpha}+\hat{\beta}X$

On passing through mean

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

59.8= $\hat{\alpha}+.35x15$

$\therefore \hat{\alpha}$ =59.8-5.25=54.55

$\therefore$ Regression equation will be –

$\hat{Y}$ =**54.55+0.35X**

On putting given values of X, the corresponding values of $\hat{Y}$ can be calculated as shown in the table.Now, we will test the hypothesis,

Suppose the null hypothesis is β =0. The formula of t is

$$t=\frac{\hat{\beta}}{\sqrt{\left(\sum e_i^2/n-2\right)}}.\sqrt{\left(\sum x_i^2\right)}\ \text{Thus,}\ t=0.35x\sqrt{\frac{754}{23.51}x3}=0.35x9.81=3.433$$

Tabulated value of't' as 3 degree of freedom is 2.353.Since tabulated value is less than calculated value of *t*, the hypothesis is to be rejected and alternative hypothesis will be accepted. Thus β is different from zero.

$$\therefore S_{\hat{\beta}} = \sqrt{\frac{\sum e^2}{(n-2).\sum x^2}} = \sqrt{\frac{23.51}{3x754}} = 0.102$$

Confidence interval at 95% level is $0.35 \pm (3.182x0.102)$,0.35 ± 0.325

Illustration:2. The following table gives the production of steel in different years at a steel factory. Find out the equation $y= a.e^{bx}$ expressing the relationship between production and year

Years	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Production (000 tons)	10.2	12.0	13.9	15.9	17.9	20.1	22.7	26.0	29.0	32.5	36.1

Solution: The given equation is $y= a.e^{bx}$ where *a* and*b* are the constant and *e* the exponential constant. Taking log to the base *e* we have-

$\log_e y = \log_e a + bx$ On putting $Y= \log_e y$ and $a_0 = \log_e a$; our equation will be

$y= a_0 + bx$ Now least squares method will be applied to estimate a_0 and b.

Year	**x**	**Production y**	**$Y = \log_e y$[1]$= log_{10}y \times 2.3025$**	**x^2**	**xY**
2011	-5	10.2	1.0086x2.3025=2.3223	25	-11.6115
2012	-4	12.0	1.0792x2.3025=2.4848	16	-9.9392
2013	-3	13.9	1.1430x2.3025=2.6318	9	-78954
2014	-2	15.9	1.2014x2.3025=2.7662	4	-5.5324
2015	-1	17.9	1.2529x2.3025=2.8848	1	-2.88848
2016	0	20.1	1.3032x2.3025=3.0006	0	0
2017	1	22.7	1.3560x2.3025=3.1222	1	3.1223
2018	2	26.0	1.4150x2.3025=3.2580	4	6.5160
2019	3	29.0	1.4624x2.3025=3.2580	9	10.1016
2020	4	32.5	1.5119x2.3025=3.4811	16	13.9244
2021	5	36.1	1.5575x2.3025=3.5861	25	17.9305
Total			**32.9015**	**110**	**13.7315**

The two normal equations are

$$\sum Y = na_0 + b\sum x \quad \text{……(i)}$$

$$\sum xY = a_0\sum x + b\sum x^2 \quad \text{……..(ii)}$$

[1]Since we are given with the log table to base of 10, to change the base to 'e' we will have to multiply the usual value of log to the base 10 with the value of '$\log_{10}$e' which is equal to 2.3025.

On putting values in the above equations

$32.9051 = 11a_0 + 0$, $\therefore$ $a_0 = 2.9914$

$13.7315 = 0 + 110b$, $b = 0.1248$

Since $a_0 = \log_e a = 2.9914$

$\therefore$ a= 19.92 and the equation is $Y = 19.92e^{.01248x}$.

The reciprocal method: The given relationship may be of the following type-

$Y = \alpha + \beta.\frac{1}{X}$ In such a situation the reciprocals of X will be taken and rest of the procedure will remain the same.

Illustration 3- Following are the observations on two variables X and Y.

X:	2	3	4	5
Y:	3.1	2.9	2.7	2.6

Estimate the equation $Y = \alpha + \beta / X$

Solution: We have to establish the relationship between Y and reciprocals of X(X*).

Y	**X**	**X***	**X*2**	**X*Y**
3.1	2	0.50	0.2500	1.550
2.9	3	0.33	0.1089	0.957
2.7	4	0.25	0.0625	0.675
2.6	5.	0.20	0.0400	0.520
11.3		**1.28**	**0.4614**	**3.702**

The estimated relationship is $Y = \hat{\alpha} + \hat{\beta}X^*$

where X* is the reciprocal of X. By formula

$$\hat{\beta} = \frac{\sum X^* Y - \frac{\sum X^* \sum Y}{n}}{\sum X^{*2} - \frac{\left(\sum X^{*2}\right)}{n}} = \frac{3.702 - \frac{1.28X11.3}{4}}{0.4614 - \frac{(1.28)^2}{4}}$$

$$= \frac{3.702 - 3.616}{0.4614 - 0.4096} = \frac{0.086}{0.0518} = 1.66$$

$\overline{Y} = \hat{\alpha} + 1.66\overline{X}^*$ or $\hat{\alpha} = \overline{Y} - 1.66\overline{X}^*$

$\hat{\alpha}$ =2.825-(1.66x.32)=2.825-.531

$\therefore$ The equation is **Y=2.294+1.66X*** **or** **Y=2.29+1.66** $\frac{1}{X}$

Illustration:4. The following statistical coefficients were deduced in the course of an examination of the relation between yield of wheat and the amount of rainfall.

Yield in Ib per acre	Annual rainfall in inches

Mean	985.0	12.8
Standard deviation	70.1	1.6
Correlation coefficient between yield and rainfall	+0.52	

Estimate the linear regression of yield on rainfall. Calculate the most likely yield of wheat per acre when the amount of rainfall is 9.2 inches.

Solution: Let yield is denoted by Y and annual rainfall is denoted by X.

Let the linear regression of yield on rainfall is $Y = \alpha + \beta X + u$

Here, Given $\overline{X}$ =12.8, $\overline{Y}$ =985, σ_v =70.1, σ_x =1.6, r=0.52

We know that $\hat{\beta}_{yx} = r.\frac{\sigma_y}{\sigma_x} = 0.52x\frac{70.1}{1.6}$ =22.78

$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 985 - 22.78x12.8 = 693.416$

Hence regression line of yield on rainfall is

$\hat{Y}$ **=693.416+22.78X**

When X=9.2 (given)then Y= 693.416+(22.78x9.2)=902.992

So when the rainfall is 9.2" then yield will be 902.992 Ibs.

Illustration: 5. A sample of 20 observations on X and Y gave the following data:-

$\sum Y = 21.9$ $\sum (Y - Y)^2 = 86.9$

$\sum X = 186.2$ $\sum (X - \overline{X})^2 = 215.4$ $\sum (X - \overline{X})(Y - \overline{Y}) = 106.4$

Answer the following:-

a) Estimate the regression of Y on X

b) Estimate the regression of X on Y

c) Compute the mean value of Y corresponding to X=10.

d) Compute the mean value of X corresponding to Y=1.5.

Solution

a) Let regression line of Y on X be

$Y = \alpha + \beta X$

we know that

$$\hat{\beta}_{yx} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{106.4}{215.4} = 0.49$$

$\overline{X} = \frac{\sum X}{n} = \frac{186.2}{20}$ =9.31 and $\overline{Y} = \frac{\sum Y}{n} = \frac{21.9}{20}$ =1.09

So, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 1.09 - (0.49x9.31) = -3.47$

Thus, estimated regression line of Y on X is

Y= -3.47+0.49X

b) Now, let the regression line of X on Y be

X= $\gamma + \delta Y$

$$\hat{\delta}_{yx} = \frac{\sum\left(X-\overline{X}\right)\left(Y-\overline{Y}\right)}{\sum\left(Y-\overline{Y}\right)^2} = \frac{106.4}{86.9} = 1.22$$

$\hat{\gamma} = \overline{X} - \hat{\delta}\overline{Y} = 9.31$ -(1.22X1.09) =7.98

Thus estimated regression line of X on Y is

X= 7.98+1.22Y

c) when X=10

then Y=-3.47+(0.49x10) =**1.43**

d) when Y=1.5

then X=7.98+(1.22x1.5)= **9.81**

Illustration 6. The following data were obtained in a sample study:-

$\sum X = 56$, $\sum Y = 40$, $\sum X^2 = 524$ $\sum Y^2 = 256$

$\sum X = 364$, N=20

Answer all of the following:

a) Estimate the regression line $Y = \alpha + \beta X$

b) Estimate the regression line $X = \gamma + \delta Y$

c) Compute the value of Y corresponding to a value 7 for X

d) Compute the value of X corresponding to a value 3 for Y

Solution:

a) Estimated regression line is $\hat{Y} = \alpha + \beta X$

$$\text{where } \hat{\beta} = \frac{\sum XY - \frac{\sum X.\sum Y}{n}}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}$$

$$= \frac{364 - \frac{56x40}{20}}{524 - \frac{56x56}{20}} = \frac{252}{367.2} = 0.686$$

$$\overline{X} = \frac{\sum X}{n} = \frac{56}{20} = 2.8, \qquad \overline{Y} = \frac{\sum Y}{n} = \frac{40}{20} = 2$$

$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 2$ -0.686x2.8=2-1.921=0.079

Thus estimated regression line becomes **Y=0.079+686X**

b) Estimated regression line is $X = \gamma + \delta Y$

$$\text{where, } \hat{\delta} = \frac{\sum XY - \frac{\sum X.\sum Y}{n}}{\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}}$$

$$= \frac{364 - \frac{56x40}{20}}{256 - \frac{40x40}{20}} = \frac{252}{256-80}$$

$$= \frac{252}{176} = 1.43$$

$$\hat{\gamma} = \overline{X} - \hat{\delta}\overline{Y} = 2.8 - 1.43x2 = 2.8 - 2.86 = -0.06$$

Now the estimated regression line becomes

X=0.06+1.43Y

c) When X=7

Then Y=0.079+.686x7=**4.881**

d) When Y=3

then X=-0.06+1.43x3=**4.23**

Illustration: 7 The following table gives ages in year of 10 husbands and their wives:-

Age of husband (X)	18	19	20	21	22	23	24	25	26	27
Age of wife (Y)	17	17	18	18	18	19	19	20	21	22

a) Estimate the linear regression of the ages of wives (Y) on the ages of husbands (X).
b) Plot the regression line on the scatter diagram.
c) Are the age of wives dependent on the ages of their husbands? Use 5% level of significance.
d) Estimate the age of the wife whose husband is 28 years old.

Solution a) Let the estimated Y on X be

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

where, $\hat{\beta} = \dfrac{\sum XY - \dfrac{\sum X \sum Y}{n}}{\sum X^2 - \dfrac{\left(\sum X\right)^2}{n}}$

and $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{X}$

To solve the $\hat{\alpha}$ and $\hat{\beta}$ we shall construct the following table-

X	Y	x-X-a (a=23)	x^2	y=Y-a (a=19)	xy	y^2
18	17	-5	25	-2	10	4
19	17	-4	16	-2	8	4
20	18	-3	9	-1	3	1
21	18	-2	4	-1	2	1
22	18	-1	1	-1	1	1
23	19	0	0	0	0	0
24	19	+1	1	0	0	0

25	20	+2	4	+1	2	1
26	21	+3	9	+2	6	4
27	22	+4	16	+3	12	9
$\sum X$ =225	$\sum Y$ =189	$\sum X$ =-5	$\sum X^2$ =85	$\sum y$ =-1	$\sum xy$ =44	$\sum y^2$ =25

$$\bar{X} = \frac{\sum X}{n} = \frac{225}{10} = 22.5, \qquad \bar{Y} = \frac{\sum Y}{n} = \frac{189}{10} = 18.9$$

Again, $\hat{\beta} = \dfrac{\sum xy - \dfrac{\sum x \sum y}{n}}{\sum x^2 - \dfrac{\left(\sum x\right)^2}{n}} = \dfrac{44 - \dfrac{(-5)x(-1)}{10}}{85 - \dfrac{(-5)^2}{10}} = \dfrac{43.5}{82.5} = 0.527$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{X}$ =18.9-(0.527x22.5) =7.044

The regression line becomes **Y= 7.044+0.527X**

d) when X=28 ,Putting the value of X in the regression line

Y=7.044+0.527x28=**218 ,**Thus when the age of husband is 28 years, the age of wife will be 22 years (approx)

(c) To test the hypothesis we shall apply the 't' test

$$t = \frac{\beta - \beta}{\sqrt{\left(\dfrac{\sum e^2}{n-2}\right)}} \; x\sqrt{\left(\sum x\right)^2}.$$

To determine '*e*' we shall construct table as following-

x	y	$\hat{Y}$	$e=Y-\hat{Y}$	e^2
18	17	16.529	0.471	0.2197
19	17	17.053	-0.053	0.0028
20	18	17.583	0.417	0.1738
21	18	18.110	-0.110	0.0121
22	18	18.637	-0.637	0.4056
23	19	19.164	-0.164	0.0268
24	19	19.691	-0.691	0.4764
25	20	20.218	-0.218	0.0475
26	21	20.745	0.255	0.0650
27	22	21.272	0.272	0.5299
				$\sum e^2$ =1.9596

Let β =0

$$t = \frac{.527}{\sqrt{\frac{1.9596}{10-2}}} x\sqrt{82.5} = \frac{.527x\sqrt{82.5x8}}{\sqrt{1.9596}}$$

$$= \frac{0.527x25.69}{1.399} = \frac{13.539}{1.399} = 9.678 \quad \therefore \text{ t calculated = 9.678}$$

t- tab at 95% confidence interval at (n-2)=10-2=8 degress of freedom =1.860

Therefore t.cal>t. tab .Since the tabulated values of t is less than its calculated value, thus the hypothesis is to be rejected i.e. $\beta \neq 0$ and alternative hypothesis will be accepted.

Illustration: 8 The following data were collected from 5 different plants in a certain industry.

Total cost (Y)	80	44	51	70	61
Production (X)	12	4	6	11	8

Answer the following questions

a) Estimate a linear total cost function $Y=\alpha+\beta X$ for the industry.

b) What is the economic significance of the estimate of α and β.

c) Estimate the total cost for a level of production of 10.

Solution: Our regression line is $Y=\alpha+\beta X$

where, $\hat{\beta} = \dfrac{\sum XY - \dfrac{\sum X \sum Y}{n}}{\sum X^2 - \dfrac{(\sum X)^2}{n}}$ and, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$

To estimate the $\hat{\alpha}$ and $\hat{\beta}$ values we shall construct the following table

Y	X	X=x-a (a=8)	X^2	Y=y-a (a=61)	XY
80	12	+4	16	+19	76
44	4	-4	16	-17	68
51	6	-2	4	-10	20
70	11	+3	9	+9	27
61	8	0	0	0	0
$\sum Y$ =306	$\sum X$ =41	$\sum X$ =+1	$\sum X^2$ =45	$\sum Y$ =+1	$\sum xy$ =191

$$\beta_{yx} = \frac{191-1/5}{45-1/5} = \frac{954}{224} = 4.25$$

$$\overline{X} = \frac{\sum X}{n} = \frac{41}{5} = 8.2 \quad \overline{Y} = \frac{\sum Y}{n} = \frac{306}{5} = 61.2$$

$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$ =61.2-4.25x8.2=26.35

Regression line of Y on X is Y=26.35+4.25Y

b) Economic significance of ' α 'and ' β '

Our estimated linear total cost equation is

Y=26.35+4.25Xwhere, α =26.35 and β =4.25

Hence, α = total fixed cost and β =marginal cost

Thus total cost =fixed cost +marginal cost x production

=total fixed cost +total variable cost.

Either production takes place or not the total fixed cost remains constant while total variable cost vary with production. The marginal cost plays an important role in the economic field. Each firm will get maximum profit where marginal cost is equal to the marginal revenue. In the field of distribution each entrepreneur should employ the factors of production to the point where marginal cost is equal to the marginal revenue product (M.R.P).In the above example, α (26.35) units) as a part of total cost is always fixed. By knowing β the entrepreneur can estimate his total of production. The accuracy of α and β is necessary for the good results

c)when X=10 , then Y=26.35+4.25x10=68.85

2.11 Let us sum up

This unit explains the simple linear regression model, its method of estimation, properties of the estimators, and goodness of fit and soled numerical problems.

2.12 Unit -End Exercises

A. Multiple Choice Questions

1. One of the following assumption is not in OLS
   A. $E(u_i / x_i)= 0$
   B. $Cov(u_i u_j / x_i x_j) = 0$
   C. No Autocorrelation
   D. No Perfect Multicollinearity
2. Find out which is not the property of a parameter?
   A. Linear in Parameter
   B. Parameters are Unbiased
   C. Parameters has the Minimum Variance
   D. Biased
3. The Co-efficient of Determination Measures
   A. The correlation between the X and Y
   B. Error
   C. Goodness of fit of the model
   D. TSS
4. The formula for testing the co-efficient with t- test is

A. Co-efficient÷ Std.Error C. Std.Error ÷ Co-efficient
B. both (A) and (C) D. larger variance ÷ Smaller Variance

5. The value of β in simple regression is
   A. B. $(X'X)^{-1} X'Y$ C. y^2 D. $(X'X)^{-1} XY$
6. The number of Explanatory variables in a simple regression is----------.
   A. Zero B.Two C. One D. More than Two
7. The property of an estimator $E(\hat{\beta}1) = \beta 1$ is termed as….
   A. Linearity B.Efficiency C.Unbiasedness D. Accuracy
8. Choose the correct one from the following about R^2
   A. R^2 = TSS/RSS C. R^2 = RSS/TSS
   B. R^2 = ESS/TSS D. R^2 = TSS/ESS

B. Short Answer and Essay type questions

1. State the assumptions of classical linear regression.
2. Derive the value of unknown parameter of α, β for the given regression equation
   $\hat{Y} = \hat{\alpha} + \hat{\beta}X_i$
3. What are the properties of estimator?
4. Prove that the estimator is linear and unbiased
5. State and prove the Gauss Markov theorem.

2.13 Reference Books

1. DamodarN.Gujarati and Sangeetha, " Basic Econometrics" Special Indian Edition , Tata McGraw Hill Education Privated Limited (Sixth Print 2010) , ISBN: 978-0-07-066005-2
2. P.G.Apte," Text book of Econometrics" Tata McGraw – Hill Publishing Company Limited
3. Dhanasekaran," Econometrics" 2nd Edition, Vrinda Publications (P) Ltd, Delhi-53 ,2011 ISBN: 978-81-8281-388-5
4. HumbertoBarreto and Frank M. Howland," Introductory Econometrics" Cambridge University Press, First South Asian Edition 2009, ISBN: 978-0-521-12358-9.
5. Dilip M. Nachane," Econometrics: Theoretical Foundations and Empirical Perspectives" Oxford University Press, Second Impression 2010 , ISBN: 978-0-19-564790-7
6. S.Shyamala,RavdeepKaur, Arul Pragasam," A text book on Econometrics: Theory and Applications" Vishal Publishing Co., Jalandhar 2017, ISBN: 81-88-646-98-9
7. S.P.Singh, Anil.K. Parashar, H.P. Singh," Econometrics and Mathematical Economics" Second Revised Editions,S.Chand and Company Ltd, New Delhi -55
8. A. Koutsoyiannis," Theory of Econometrics" Second Edition Palgrave – New York,2004,ISBN: 0-333-77822-7
9. Maddala .G.S.(1997), "Econometrics" , McGraw Hill, New York.
10. Johnston. (1997)," Econometric Methods" McGraw Hill, 4th Edition, New Delhi.

UNIT3

MULTIPLE LINEAR REGRESSION MODEL

3.1 Objectives

The specific objective of this chapter intends towards the student are as follows

- To make them with familiarity of multiple regression and its properties
- To understand the multiple regressions for using in reality as a researcher, as a field surveyor, and desk researcher.

3. 2 Introduction

This unit made an attempt to explain the Multiple Regression Model and its properties. Any economic activity is not with single factor, but determined by more than one factors. The daily routine life demand in early morning tea or coffee determined the factors like input cost of its preparation, taste and preference, and its competitive beverages prices. Every economic activity determined by more than one variable from micro to macro. Hence it is essential to study this chapter for dealing such a kind of situation**.**

3.3 Meaning of Multiple regression

Multiple regression models are an extended form of more than one explanatory variable. The multiple regression models are designed to describe economic relationships as an extension of the simple regression model.

Let us start with the theoretical proposition that changes in one variable can be explained by change in several other variables. Such a relationship is described in simple way by a multiple linear regression equation of the form

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i \qquad ..... (1)$$

where Y denotes the dependent variable, the X's are explanatory variables and U is a stochastic disturbance term.

3.4 The assumptions underlying the Method of Ordinary Least Squares

Basic assumption which are as follow

a. u_i is normally distributed
b. $E(u_i)=0$
c. $E(u_i^2)=\sigma_u^2$
d. $E(u_i u_j)= 0 \qquad$ for $i \neq j$
e. Each of the explanatory variables is non-stochastic with fixed value in repeated samples and such that for any sample size $\sum_{i=1}^{n}\left(X_{ki} - \overline{X}_k\right)^2 / n$ is a finite number different from zero.
f. The number of observations exceeds the number of coefficients to be estimated.
g. No exact linear relation exists between any of the explanatory (X's) variables.

3.5 Estimation of Model by Method of Ordinary Least Square

Ordinary Least squares principle applies over the regressions model which is expressed in matrix. As the regression model was

$Y= X\beta + u$ where

$$y = \begin{bmatrix} y1 \\ y2 \\ y3 \\ . \\ . \\ yn \end{bmatrix}, X = \begin{bmatrix} x_{21}x_{31}...x_{k1} \\ x_{22}x_{32}...x_{k2} \\ \\ . \\ . \\ x_{2n}x_{3n}...x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_2 \\ \beta_3 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}. u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ u_n \end{bmatrix}$$

Sum of the squared residuals is

$$\sum u_i^2 = \sum_{i=1}^{n} e^2 i = e'e = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ e_n \end{bmatrix} \begin{bmatrix} e_1 e_2 ... e_n \end{bmatrix} = e_1^2 + e_2^2 + ... + e_n^2$$

$$= \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right)$$

$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}X'X\hat{\beta}$$

(Since $\beta'X'Y$ is a scalar, thus it will equal to $Y'X\hat{\beta}$)

$$\therefore \quad \sum e_i^2 = Y'Y - 2\hat{\beta}X'Y + \hat{\beta}'X'X\hat{\beta}$$

For least squares first differentiation should be equal to zero

or $$\frac{\partial\left(\sum e_i^2\right)}{\partial\hat{\beta}} = 2X'Y + 2X'X\hat{\beta}$$

$$\therefore 2X'X\hat{\beta} - 2X'Y = 0$$

$$(X'X)\hat{\beta} = X'Y$$

Premultiplying by $(X'X)^{-1}$ to both the sides we have,

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'Y$$

$$\therefore \quad \hat{\beta} = (X'X)^{-1}X'Y \text{ where } X' \text{ is column vector, X=row vector.}$$

This is the fundamental result for the least squares estimators. To determine the β_0 we shall use $\overline{Y} - \hat{\beta}_1\overline{X}_1 - \hat{\beta}_2\overline{X}_2 - \ldots - \hat{\beta}_k\overline{X}_k = \hat{\beta}_0$

3.6 Properties of Ordinary Least Squares Estimators

Least squares estimators are best linear and unbiased (BLUE)

Let us take linear model be- $Y=X\beta+u$

3.6.1 i) The Least Square Estimators are Unbiased Estimators-

Since $\hat{\beta} = (X'X)^{-1}X'Y$

$Y=X\beta+u$

$$\therefore \quad \hat{\beta} = (X'X)^{-1}X'(X\beta + u)$$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u$$

$$= \beta + (X'X)^{-1}X'u$$

taking expectations

$$E(\hat{\beta}) = \beta + (X'X)^{-1}X'E(u)$$

Since E(u)=0

$$\therefore E(\hat{\beta}) = \beta$$

Thus are unbiased estimators of

3.6.2 ii) The Least Square Estimators are linear estimators- Since least squares estimators have linear relation with Y, so they are linear.

3.6.3 iii) The Least Square Estimators are best estimators-Now we will prove that our estimators are best among all the estimators, since

$$\hat{\beta} = \beta(X'X)^{-1}X'u \qquad \therefore \quad (\hat{\beta} - \beta) = (X'X)^{-1}X'u$$

$$\text{Var.}(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$$

$$= E\left[\{(X'X)^{-1}X'u\}\{(X'X)^{-1}X'u\}'\right]$$

$$= E\left[(X'X)^{-1}X'u.u'X(X'X)^{-1}\right]$$

Since $\left[(X'X)^{-1}\right] = (X'X)^{-1}$

$$= (X'X)^{-1}X'.E(uu').X(X'X)^{-1}$$

$$= (X'X)^{-1}X'.\sigma_u^2.I_nX(X'X)^{-1}$$

$$= \sigma_u^2.I_n.(X'X)^{-1}.X'X(X'X)^{-1} = \sigma_u^2(X'X)^{-1}$$

$\therefore \text{Var}(\hat{\beta}) \leq$ var (u). Hence $\hat{\beta}$ is an best estimator.

Now we shall prove the more general result which is a special case. The more general result also has its applications in predication problem. Let us consider a relation

b=(A+B)Y where $Y=X\beta+u$, $A=(X'X)^{-1}X'$ and B is a constant.

or b=(A+B) $(X\beta+u)$

$=(A+B)X\beta+(A+B)u$

Taking expectation of both sides we have

E(b) $= (A+B)X\beta$ since E(u)=0

$=AX\beta+BX\beta$

$=\beta+BX\beta$ since $AX=I_n$

E(b) $=\beta$ only if BX=0.

Thus b is an unbiased estimator.

Again $b=\beta+(A+B)u$

$b-\beta=(A+B)u$

and $(b-\beta)' = \{(A+B)u\}' = u'(A+B)'$

$$E\{(b-\beta)\{b-\beta'\}\} = E\{(A+B)uu'(A+B)'\}$$

$$= E(uu')(A+B)(A'+B')$$

$$= \sigma_u^2(A+B)(A'+B')$$

$$= \sigma_u^2\{AA'+BA'+AB'+BB'\}$$

$$= \sigma_u^2 \left\{ (X'X)^{-1} X'X (X'X)^{-1} + BX(X'X)^{-1} + (X'X)^{-1} X'B' + BB' \right\}$$

Since $(X'X)^{-1}(X'X) = I_n; BX = 0$

$$\therefore \text{Var.(b)} = \sigma_n^2 \left\{ (X'X)^{-1} + BB' \right\}$$

$$= \sigma_n^2 (X'X)^{-1} + \sigma_n^2 BB'$$

$$= \text{Var.}\hat{\beta} + \sigma_u^2 BB'$$

$$\therefore \text{var (b) } - \text{var}(\hat{\beta}) \geq 0$$

Let us consider the vector Ci with unity in the ith position and zero elsewhere, then var (b) $\geq$ Var (β)

Thus other estimators will either have greater or at least equal values to least squares estimator. So we can say that only least squares estimators have smallest variance among all linear unbiased estimators. Hence our least squares estimators are the best linear and unbiased (BLUE).

3.7 R^2 and the adjusted R^2 (Goodness of Fit)

R^2 is a non-decreasing function of the number of explanatory variables or regressors present in a model; as the number of regressors increases, R^2 almost invariably increases and never decreases. Stated differently, an additional X variable will not decrease R^2. It explains the goodness of fit of a model.

$R^2 = \frac{ESS}{TSS}$ Where ESS stands for Explained Sum Square and TSS stands for Total Sum Square.

$$= 1 - \frac{RSS}{TSS}$$

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

Now $\sum y_i^2$ is independent of the number of X variables in the model because it is simply $\sum (Y_i - \overline{Y})^2$. The RSS, $\sum \hat{u}_i^2$, however, depends on the number of regressors present in the model. Intuitively, it is clear that as the number of X variables increases, $\sum \hat{u}_i^2$ is likely to decrease (at least it will not increase); hence R^2 as defined it will increase. In view of this, in comparing two regression models with the same dependent variable but differing number of X variables, one should be very wary of choosing the model with the highest R^2.

To compare two R^2 terms, one must take into account the number of X variables present in the model. This can be done readily if we consider an alternative coefficient of determination, which is as follows:

$\overline{R}^2 = 1 - \frac{\sum \hat{u}_i^2/(n-k)}{\sum y_i^2/(n-1)}$ where k = the number of parameters in the model including the intercept term. (In the three-variable regression, k = 3. Why?) The R^2 thus defined is known as the adjusted R^2, denoted by $\overline{R}^2$. The term adjusted means adjusted for the degrees of freedom associated with the sums of squares entering into: $\sum \hat{u}_i^2$ has n-k degrees of freedom in a model involving k parameters, which include the intercept term, and $\sum y_i^2$ has n-1 degree of freedom. (Why?) For the three-variable case, we know that $\sum \hat{u}_i^2$ has n-3 degrees of freedom. $\overline{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_Y^2}$ where $\hat{\sigma}^2$ is the residual variance, an unbiased estimator of true σ^2, and S_Y^2 is the sample variance of Y. It is easy to see that $\overline{R}^2$ and R^2 are related because, substituting the value of R^2, obtain $\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$. It implies that as the number of X variables increases, the adjusted R^2 increases less that the unadjusted R^2; and $\overline{R}^2$ can be negative, although R^2 is necessarily nonnegative. In case $\overline{R}^2$ turns out to be negative in an application, its value is taken as zero. Which R^2should one use in practice? As Theil notes: ….it is good practice to use $\overline{R}^2$ rather than R^2 because R^2 tends to give an overly optimistic picture of the fit of the regression, particularly when the number of explanatory variables is not very small compared with the number of observations, explanatory variables is not very small compared with the number of observations.

3.8 Solved Numerical Problems

Illustration: 1. A random sample of five families yields the following data

Family	**A**	**B**	**C**	**D**	**E**
Saving S(in hundred Rs.)	6	12	10	7	3
Income Y (in thousand Rs)	8	11	9	6	6
No. of children, N	5	2	1	3	4

Estimate the regression line of S on Y and N

Solution: The linear regression line will be

$$S = \beta_0 + \beta_1 Y + \beta_2 N + u$$

So estimated line will be

$$\hat{S} = \hat{\beta}_0 + \hat{\beta}_1 Y + \hat{\beta}_2 N$$

To apply least square method, the three normal equation are

$$\sum S = n.\hat{\beta}_0 + \hat{\beta}_1 \sum Y + \hat{\beta}_2 \sum N$$

$$\sum SY = \hat{\beta}_0 . \sum Y + \hat{\beta}_1 \sum Y^2 + \hat{\beta}_2 \sum NY$$

$$\sum SN = \hat{\beta}_0 . \sum N + \hat{\beta}_1 \sum NY + \hat{\beta}_2 \sum N^2$$

Values of these equations will be calculated as follows:-

Family	S	Y	N	Y^2	N^2	SY	SN	NY
A	6	8	5	64	25	48	30	40
B	12	11	2	121	4	132	24	22
C	10	9	1	81	1	90	10	9
D	7	6	3	36	9	42	21	18
E	3	6	4	36	16	18	12	24
Total	**38**	**40**	**15**	**338**	**55**	**330**	**97**	**113**

On putting values in normal equations,

$38 = 5\hat{\beta}_0 + 40\hat{\beta}_1 + 15\hat{\beta}_2$ …….(1)

$330 = 40\hat{\beta}_0 + 338\hat{\beta}_1 + 113\hat{\beta}_2$ …… (2)

$97 = 15\hat{\beta}_0 + 113\hat{\beta}_1 + 55\hat{\beta}_2$ ……....(3)

On multiplying equation (1) by 8 and subtracting it from equation (2)

$18\hat{\beta}_1 + 7\hat{\beta}_2 = 26$ ……(4)

On multiplying equation (1) by 3 and substracting it from equation (3)

$7\hat{\beta}_1 + 10\hat{\beta}_2 = 17$ ……(5)

On multiplying equation (4) by 10 and (5) by 7

$180\hat{\beta}_1 + 70\hat{\beta} = 260$

$$\frac{+49\hat{\beta}_1 - 70\hat{\beta}_2 = +119}{131\hat{\beta}_1 = 141}$$ on subtracting

or $\hat{\beta}_1 = \frac{141}{131} = 1.076$

On putting the value of $\hat{\beta}_1$ in equation (5)

$7x1.076-10\hat{\beta}_2 = 17$

$-10\hat{\beta}_2 = 9.466$ or $\hat{\beta}_2 = -.9466 = -.947$

Now the equation will pass through mean values. So , $\bar{S} - \hat{\beta}_0 + \hat{\beta}_1\bar{Y} + \hat{\beta}_2\bar{N}$

∴ $7.6\,\hat{\beta}_0 + (8x1.076) + 3(-947) = \hat{\beta}_0 = 1.833$

∴ Our estimated relation will be

S=1.833+1.076Y-.947N

Illustration: 2 The following matrix gives the variances and covariances of three variables:-

X_1=log food consumption per capita

X_2=log food price

X_3=log disposable income per capita

$$\begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{bmatrix} 7.59 & 3.12 & 26.99 \\ & 29.16 & 30.80 \\ & & 133.0 \end{bmatrix} \end{array}$$

On the assumption that the demand relationship may be adequately represented by a function of the firm $Y_1 = AY_2^{\alpha} Y_3^{\beta}$ (where $X_i = \log Y_i$) estimate the income elasticity of demand.

Solution: From the above matrix we have,

$X_1^2 = 7.59$ $X_1X_2 = 3.12$ $X_1X_3 = 26.99$

$X_2^2 = 29.16$ $X_2X_3 = 30.80$

$X_3^2 = 133.0$

Given regression line is

$$Y_1 = AY_2^{\alpha}Y_3^{\beta} \quad \ldots(1)$$

Taking log of both sides we have,

$$\log Y_1 = \log A + \alpha \log Y_2 + \beta \log Y_3 .$$

or $$X_1 = \alpha_0 + X_2 + X_3 . \quad \ldots(2)$$

where $X_i = \log Y_i$ (given) and $\log A = \alpha_0$

From the regression line we have

$$\alpha_1 = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \qquad X = [X_2 \; X_3]$$

$$XX = \begin{bmatrix} X_2 \\ X_3 \end{bmatrix}[X_2X_3] = \begin{bmatrix} X_2^2 & X_2X_3 \\ X_2X_3 & X_3^2 \end{bmatrix} = \begin{bmatrix} 29.16 & 30.80 \\ 30.80 & 133.0 \end{bmatrix}$$

Since we know,

$$(X'X)^{-1} = \frac{\text{Adj.of}(X'X)}{|X'X|}$$

$$|X'X| = \begin{vmatrix} 29.16 & 30.80 \\ 30.80 & 133.0 \end{vmatrix} (133X29.16 - 30.80X30.80)$$

$$3878.28 - 94.64 = 2929.64$$

$$\text{Transpose of } X'X = \begin{bmatrix} 29.16 & 30.80 \\ 30.80 & 133.0 \end{bmatrix}$$

Cofactor of 29.16 =133

Cofactor of30.80 =30.80 {Using $(-1)^{i+j}$where i=no of rows, j=no of columns}

Cofactor of 30.80 =30.80

Cofactor of 133 =29.66

$$\text{Adjoint of } X'X = \begin{bmatrix} 133.0 & -30.80 \\ -30.80 & 29.16 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{2929.64}\begin{bmatrix} 133.0 & -30.80 \\ -30.80 & 29.16 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{133}{2929.64} & \frac{-30.80}{2929.64} \\ \frac{-30.80}{2929.64} & \frac{29.16}{2929.64} \end{bmatrix} = \begin{bmatrix} .045 & -.010 \\ -.010 & .009 \end{bmatrix}$$

$$\text{Now } X'X_1 = \begin{bmatrix} X_2 \\ X_3 \end{bmatrix} X_1 = \begin{bmatrix} X_1X_2 \\ X_1X_3 \end{bmatrix} = \begin{bmatrix} 3.12 \\ 26.99 \end{bmatrix}$$

$$\hat{\alpha} = (X'X)^{-1}X'X_1 = \begin{bmatrix} .045 & -.010 \\ -.010 & .009 \end{bmatrix} \begin{bmatrix} 3.12 \\ 26.99 \end{bmatrix}$$

$$= \begin{bmatrix} 3.12X.045 - 26.99X.010 \\ -3.12X.010 + 26.99X.009 \end{bmatrix} = \begin{bmatrix} 0.1404 - 0.2699 \\ -0.0312 + 0.2429 \end{bmatrix}$$

$$= \begin{bmatrix} -0.1295 \\ 0.2117 \end{bmatrix} \text{Thus } \alpha = 0.1295 \text{ and } \beta = 0.2177$$

Our given equation is, $X_1 = \alpha_0 + \alpha X_2 + \beta X_3$ .where we have estimated α **=0.1295** and β **=0.2117**

Here β =0.2117 is known as income elasticity of demand.

(since there are some structural relationships, which describe the behaviour of the individuals in the economy. These are, for instance demand function, production function and supply function. These structural relationships also involve structural parameters, which are to be estimated by statistical methods. Examples of such parameters are elasticity of demand with respect to price, elasticity of demand with respect to income, marginal propensity to consume and marginal production.)

Illustration: 3 Three related variates X_1, X_2, X_3 take the following sets of values:-

X1	1	2	3	4	5
X2	2	1	5	4	3
X3	3	1	4	5	2

a) Show that the regression plan of X_1 on X_2 and X_3 is

$18X_1-17X_2+10X_3=33$

b) Also test the null hypothesis $H_0(\beta_2=0)$ against alternative hypothesis $H_1(\beta_2 \neq 0)$ at 5% level of significance.

Solution: Let the regression line be

$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3$.

We have $\hat{\beta} = (X'X)^{-1}X'X_1$ where, $\hat{\beta} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix}$, $X = [X_2X_3]$ and $X' = \begin{bmatrix} X_2 \\ X_3 \end{bmatrix}$

We shall estimate $\hat{\beta}$ as following

X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
1	2	3	1	4	9	2	3	6
2	1	1	4	1	1	2	2	1
3	5	4	9	25	16	15	12	20
4	4	5	16	16	25	16	20	20
5	3	2	25	9	4	15	10	6
15	15	15	55	5	55	50	47	53

$$\overline{X}_1 = \frac{\sum X_1}{n} = \frac{15}{5} = 3 \qquad \overline{X}_2 = 3, \overline{X}_3 = 3$$

Now we shall construct the following quantities in terms of deviation around the means

$$\sum x_2^2 = \sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n} = 55 - \frac{15X15}{5} = 10$$

$$\sum x_3^2 = 55 - \frac{15X15}{5} = 10, \sum x_1^2 = 10$$

$$\sum x_1x_2 = \sum X_1X_2 - \frac{\sum X_1 \sum X_2}{n} = 50 - \frac{15x15}{5} = 5$$

$$\sum x_1x_3 = 47 - 45 = 2, \sum x_2x_3 = 53 - 45 = 8$$

$$X'X = \begin{bmatrix} X_2 \\ X_3 \end{bmatrix}[X_2X_3] = \begin{bmatrix} x_2^2 & x_2x_3 \\ x_2x_3 & x_3 \end{bmatrix} = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$$

$$|X'X| = \begin{vmatrix} 10 & 8 \\ 8 & 10 \end{vmatrix} = (100 - 64) = 36$$

Cofactor of 10=10, cofactor of 8=-8, cofactor of 8=-8, Cofactor of 10=10.

$$(X'X)^{-1} = \frac{\text{Adjoint of}(X'X)}{\text{Determinant of}(X'X)}, \text{Adjoint of } (X'X) = \begin{bmatrix} 10 & -8 \\ -8 & 10 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{36}\begin{bmatrix} 10 & -8 \\ -8 & 10 \end{bmatrix} = \begin{bmatrix} 0.278 & -0.222 \\ -0.222 & 0.278 \end{bmatrix}$$

$$X'X_1 = \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}x_1 = \begin{bmatrix} x_2x_1 \\ x_3x_1 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$(X'X)^{-1}XX_1 = \begin{bmatrix} 0.278 & -0.222 \\ -0.222 & 0.278 \end{bmatrix}\begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.278x5 - 0.222x2 \\ -0.222x5 + 0.278x2 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'X_1 = \begin{bmatrix} 0.946 \\ -0.554 \end{bmatrix} = \begin{bmatrix} \beta_2 \\ \beta_3 \end{bmatrix} \qquad \therefore \qquad \hat{\beta}_2 = 0.946, \qquad \hat{\beta}_2 = -0.554$$

$$\overline{X}_1 - \hat{\beta}_2\overline{X}_2 - \hat{\beta}_3\overline{X}_3 = \hat{\beta}_1$$

$$3 - .946X3 + .554X3 = \hat{\beta}_1$$

$$3 - .946X3 + .554X3 = \hat{\beta}_1$$

$$\therefore \qquad \hat{\beta}_1 = 1.824$$

Regression line of X1 on X2 and X3 is

$X_1 = 1.824 + 0.946X_2 - 0.554X_3$.

or we can write it in another way; on multiplying both the sides of eq. by18

$18X_1 = 32.83 + 17.028X_2 - 9.97X_3$.

or $18X_1 = 33 + 17X_2 - 10X_3$

or $18X_1 = 17X_2 + 10X_3 = 33$. (proved)

b) Test of significance

1	$\hat{X}_1$	$e = (X_1 - \hat{X}_1)$	e^2
1	2.054	-1.054	1.110
2	2.216	-0.216	0.0467
3	4.338	-1.338	1.790
4	1.838	+2.162	4.674
5	3.554	+1.446	2.091
			$\sum e^2 = 7.875$

Applying the 't' test- $t = \dfrac{\hat{\beta}_2 - \beta_2}{\sqrt{\dfrac{\sum e^2}{n-k}}}\sqrt{a_{ii}}$ β_2 =hypothetical parameter , e= error term,

k= no.ofparameters , a_{ii}= ith diagonal element in $(X'X)^{-1}$

Putting the values in the above formula

$$t = \frac{0.946}{\sqrt{\dfrac{7.875}{5-2}}} X \sqrt{0.278} = \quad \frac{0.946}{0.854} = 1.108 \text{ degrees of freedom 5-3=2}$$

t- tab at 5% level of significance for 2 degrees of freedom =2.920 and t.cal.=1.108

$\therefore$ t-cal.< t-tab. Thus accept the hypothesis $\therefore \beta_2 = 0$. Thus we shall accept the null hypothesis $H_0(\beta_2 = 0)$ i.e There is no relationship between X_1 and X_2.

Illustration:4 The following table shows the weights (X_1) to the nearest pound, heights (X_2) to the nearest inch and ages (X_3) to the nearest year of 12 boys:-

Weight (X1)	Height (X2)	Age (X3)
64	57	8
71	59	10
53	49	6
67	62	11
55	51	8
58	50	7
77	55	10
57	48	9
56	52	10
51	42	6
76	61	12
68	57	9

Estimate the least squares regression line to predict the weight of a boy of given height and age.

Solution: Let the regression line of X_1on X_2 and X_3 be $X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3$

To determine the values of parameters I e. β_1 β_2 and β_3 we shall construct the following table-

X_1	X_2	X_3	X_1X_2	$_1X_2$	X_2X_3	X_1^2	X_2^2	X_3^2
64	57	8	3648	512	456	4096	3249	64
71	59	10	4189	710	590	5041	3481	100
53	49	6	2597	318	294	2809	2401	36
67	62	11	4145	737	682	4489	3844	121
55	51	8	2850	440	408	3025	2601	64
58	50	7	2900	406	350	3364	2500	49
77	55	10	4235	770	550	5929	3025	100
57	48	9	2736	513	432	3249	2304	81
56	52	10	2912	560	520	3136	2704	100
51	42	6	2142	306	252	2601	1764	36
76	61	12	4636	912	732	5776	3721	144
68	57	9	3876	612	5133	4624	3249	81
753	643	106	40830	6796	5779	48139	34843	976

$$\overline{X}_1 = \frac{753}{12} = 62.75, \overline{X}_2 = \frac{643}{15} = 53.58, \overline{X}_3 = \frac{106}{12} = 8.83$$

Setting of value to the actual mean-

$$\sum x_2^2 = \sum x_2^2 - \left(\sum X_2\right)^2 / n = 34843 - (643)^2 / 12 = 388.92$$

$$\sum x_3^2 = \sum x_3^2 - \left(\sum X_3\right)^2 / n = 976 - (106)^2 / 12 = 39.67$$

$$\sum x_1 x_2 = \sum x_1 x_2 - \left(\sum X_1 . \sum X_2\right) / n = 40830 - (753X643) / 12 = 481.75$$

$$\sum x_1 x_3 = \sum x_1 x_3 - \left(\sum X_1 . \sum X_3\right) / n = 6796 - (753X106) / 12 = 144.5$$

$$\sum x_2 x_3 = \sum x_2 x_3 - \left(\sum X_2 . \sum X_3\right) / n = 5779 - (643X106) / 12 = 99.17$$

on applying the formula

$$\hat{\beta}_2 = \frac{\left(\sum x_1 x_2 . \sum x_3^2\right) - \left(\sum x_1 x_3 . \sum x_2 x_3\right)}{\sum x_2^2 . \sum x_3^2 - \left(\sum x_2 x_3\right)^2} = \frac{(481.75x39.67) - (144.5 - 99.17)}{(388.92x39.67) - (99.17)^2}$$

$$= \frac{19111.02 - 14330.06}{15428.46 - 9834.69} = \frac{4780.96}{5593.77} = 0.85$$

$$\hat{\beta}_3 = \frac{\left(\sum x_1 x_3 . \sum x_2^2\right) - \left(\sum x_1 x_2 . \sum x_2 x_3\right)}{\sum x_2^2 . \sum x_3^2 - \left(\sum x_2 x_3\right)^2} = \frac{(44.5x388.92) - (481.75x99.17)}{(388.92x39.67) - (99.17)^2}$$

$$= \frac{56198.94 - 47775.15}{15428.46 - 9834.69} = \frac{8423.79}{5593.77} = 1.51$$

Now $\hat{\beta}_1 = \overline{X}_1 - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3$

=62.75-(0.85X53.58)-(1.51X8.83)

=62.75-45.54+13.33=3.88 Thus regression line is $X_1=3.88+0.85X_2+1.51X_3$.

Illustration: 5 From the following data compute the regression line of X_1 on X_2 and $_3$.

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
X_1	100	106	107	120	110	116	123	133	137
X_2	100	104	106	111	111	115	120	124	126
X_3	100	99	110	126	113	103	102	103	98

Where X_1= Index of imports of goods and services to U.S.A at constant (2000) prices.

X_2=Index of gross U.S.A product at 2000 prices.

X_3= Ratio of indices of prices of imports and general U.S.A output respectively.

Solution: Let the estimated regression line of X_1 onX_2 and X_3 be

$X_1 = \beta_1 + \beta_2 X_2 + \beta_3 X_3$ From the above table firstly we shall compute the mean

n=9 $\sum X_1$ =1052, $\sum X_2$ =1017, $\sum X_3$ =954

$$\overline{X}_1 = \frac{\sum \overline{X}}{n} = \frac{1052}{9} = 116.9; \overline{X}_2 = \frac{1017}{9} = 113; \overline{X}_3 = \frac{954}{9} = 106$$

$\sum X_1X_2$ =119,750, X_1X_3=111.433 $\sum X_2X_3$ =107, 690

$\sum X_1^2$ =124,288 $\sum X_2^2$ =115,571, $\sum X_3^2$ =101,772

Now we shall compute in terms of deviation from actual mean

$$\sum x_1x_2 = x_1x_2 - \frac{\sum X_1 . \sum X_2}{n} = 119750 - \frac{1052x1017}{9} = 874$$

$$\sum x_1x_3 = x_1x_3 - \frac{\sum X_1 . \sum X_3}{n} = 111433 - \frac{1052x954}{9} - 79$$

$$\sum x_2x_3 = \sum X_2X_3 - \frac{\sum X_2 . \sum X_3}{n} = 107690 - \frac{1017x954}{9} = -112$$

$$\sum x_1^2 = \sum x_1^2 - \frac{\left(\sum X_1\right)^2}{n} = 124228 - \frac{1052x1052}{9} = 126089$$

$$\sum x_2^2 = \sum x_2^2 - \frac{\left(\sum X_2\right)^2}{n} = 115571 - \frac{1017x1017}{9} = 650$$

$$\sum x_3^2 = \sum x_3^2 - \frac{\left(\sum X_3\right)^2}{n} = 101772 - \frac{954x954}{9} = 648$$

We know that $\hat{\beta} = (XX')^{-1}X'X_1$ where, X= [x_2 x_3] and $X' = \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}$

Now $X'X = \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}[x_2 \quad x_3] = \begin{bmatrix} x_2^2 & x_2x_3 \\ x_2x_3 & x_3^2 \end{bmatrix} = \begin{bmatrix} 650 & -112 \\ -112 & 648 \end{bmatrix}$

$$|X'X| = \begin{vmatrix} 650 & -112 \\ -112 & 648 \end{vmatrix} = (650X648 - 112X112) = 408656$$

Cofactor of 650=648 since sign for cofactor is $(-1)^{i+j}$

Cofactor of -112=112 where i=no. of rows,

Cofactor of -112=112 j=no. of columns

Cofactor of 648=650

Adjoint of $(X''X) = \begin{bmatrix} 648 & 112 \\ 112 & 650 \end{bmatrix}$ Since, Inverse $= \frac{\text{Adjoint}}{\text{Determinant}}$

$$\therefore (X'X)^{-1} = \frac{1}{408656}\begin{bmatrix} 648 & 112 \\ 112 & 650 \end{bmatrix} = \begin{bmatrix} \frac{648}{408656} & \frac{112}{408656} \\ \frac{112}{408656} & \frac{650}{408656} \end{bmatrix}$$

(Since λ [A]=[λ A]where, A is a matrix and λ is scalar)

$$(X'X)^{-1} = \begin{bmatrix} 0.00158 & 0.00027 \\ 0.00027 & 0.00159 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} x_2 \\ x_3 \end{bmatrix}X_1 = \begin{bmatrix} x_1x_2 \\ x_1x_3 \end{bmatrix} = \begin{bmatrix} 874 \\ -79 \end{bmatrix}$$

$$\therefore \quad \hat{\beta} = (X'X)^{-1}X'X_1 = \begin{bmatrix} 0.00158 & 0.00027 \\ 0.00027 & 0.000159 \end{bmatrix}\begin{bmatrix} 874 \\ -79 \end{bmatrix}$$

$$= \begin{bmatrix} 0.00158X874 + 0.00027X - 79 \\ 0.00027X874 + 0.00159X - 79 \end{bmatrix} = \begin{bmatrix} 1.364 \\ 0.114 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} \qquad \therefore \hat{\beta}_2 = 1.364 \quad \hat{\beta}_3 = 0.113$$

Now $\hat{\beta}_1 = \overline{X}_1 - \hat{\beta}_2\overline{X}_2 - \hat{\beta}_3\overline{X}_3$ =116.9-1.364x113-0.114x106=-49.33

$\therefore$ The estimated regression line becomes $X1=-49.33+1.364X_2+0.114X_2$.

3.9. Let us sum up

If we have several repressors in a regression model, how do we estimate and what are the assumptions of that model, properties of estimators such as linear, unbiased, minimum variance, explained followed by goodness of fit of model with solved numerical problems.

3.10 Unit -End Exercises

A. Multiple Choice Questions

1. The term Multiple regression Stands for
   A. Regressing more than one explanatory variables
   B. Regressing no variables
   C. Many regression
   D. Regressing one explanatory variable
2. An estimator is consistent if..?
   A. It converges to the true value as the sample size remains same.
   B. It converges to the true value as the sample size gets smaller.
   C. It converges to the true value as the sample size gets larger.
   D. Above all.
3. The value of $\widehat{\beta 0}$ is

   A. $\hat{\beta}_0 = \bar{y} - \beta_1\bar{x}$ B. $\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$ C. . Zero D,. One
4. Regression model in which more than one independent variable is used to predict the dependent variable is called
   A. a simple linear regression model C. a multiple regression model
   B. an independent model D. none of the above
5. The Value of β estimator is
   A. $\hat{\beta} = (X'X)^{-1}X'Y$
   B. $\hat{\beta} = (X'X)^{-1}X'X$
   C. $\hat{\beta} = (X'Y)^{-1}X'X$
   D. $\hat{\beta} = (X'Y)^{-1}Y'Y$

B. Short Answer and Essay type Questions

1. What is Multiple Regression?
2. Describe the method of estimation of unknown parameter with matrix.
3. Derive the β value in terms of matrix as $(X'X)^{-1}X'Y$
4. Enumerate the properties of linear, unbiased and minimum variance of estimator.
5. What do meant by goodness of fit? Explain it.

3.11. Reference Books

1. DamodarN.Gujarati and Sangeetha, " Basic Econometrics" Special Indian Edition , Tata McGraw Hill Education Privated Limited (Sixth Print 2010) , ISBN: 978-0-07-066005-2
2. P.G.Apte," Text book of Econometrics" Tata McGraw – Hill Publishing Company Limited.
3. Dhanasekaran," Econometrics" 2nd Edition, Vrinda Publications (P) Ltd, Delhi-53 ,2011 ISBN: 978-81-8281-388-5
4. HumbertoBarreto and Frank M. Howland," Introductory Econometrics" Cambridge University Press, First South Asian Edition 2009, ISBN: 978-0-521-12358-9.
5. Dilip M. Nachane," Econometrics: Theoretical Foundations and Empirical Perspectives" Oxford University Press, Second Impression 2010 , ISBN: 978-0-19-564790-7
6. S.Shyamala,RavdeepKaur, Arul Pragasam," A text book on Econometrics: Theory and Applications" Vishal Publishing Co., Jalandhar 2017, ISBN: 81-88-646-98-9
7. S.P.Singh, Anil.K. Parashar, H.P. Singh," Econometrics and Mathematical Economics" Second Revised Editions,S.Chand and Company Ltd, New Delhi -55
8. A. Koutsoyiannis," Theory of Econometrics" Second Edition Palgrave – New York,2004,ISBN: 0-333-77822-7
9. Maddala .G.S.(1997), "Econometrics" , McGraw Hill, New York.
10. Johnston. (1997)," Econometric Methods" McGraw Hill, 4th Edition, New Delhi.

UNIT 4

PROBLEMS OF SINGLE EQUATION MODELS

Structure

4.1 Objectives

After going through the unit you will be able to:

- Understand the concepts and issues of violation of assumptions
- Sense the causes and consequences of violation of assumptions.
- Know the method of detection and remedies for violation of assumptions.

4.2 Introduction

The econometric models are constructed by introducing the random variable 'U_i' for to take into account of influence of various errors, such as (a) errors of omitted variable (b) errors of the mathematical form of the model (c) errors of measurement of the dependent variable and (d) the effects of the erratic element which is inherent in human behaviour.

We studied the role of random variable 'U_i' and the reason for it is introduced into model in unit one and in unit two, under assumptions. Further to get valid and representative results one must be familiar with the expected consequences from non-fulfillment of an assumption. Hence, this chapter, we will discuss the causes, consequences, detection and remedies which are to be made if any one of the basic assumption is violated.

4.3 Violations of Assumptions

There are ten assumptions framed for execution of classical linear regression model under Ordinary Least Square method of estimation. In that the assumptions related with multicollineariy, Heteroscedasticity, Autocorrelation and Specification error are important one for getting the consistent and efficient parameters of intercept(α) Slope(β) and stochastic error term(U_i.

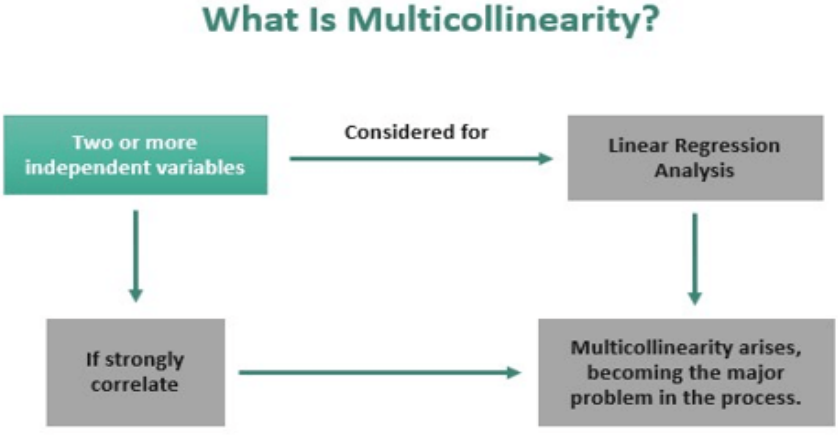
4.4.Multicollinearity

The classical linear regression model (CLRM) assumes that there is no multicollinearity among the regressors included in the regression model.

4.4.1. Meaning:

Multicollinearity refers to the existence of more than one exact linear relationship, and collinearity refers to the existence of a single linear relationship. Originally, Multicollinearity meant that the existence of a "perfect" or exact, linear relationship among some or all explanatory variables of a regression model.

In Classical linear regression model assume that there is no multicollinearity, among the explanatory variables (Xs). The reasoning is this: if multicollinearity is perfect. Then the regression coefficients of explanatory variables are in determinate and their standard errors are determinate and their standard errors are infinite. If multicollinearity is less than perfect, the regression coefficients possess large standard errors, which means the coefficients cannot be estimated with great precision of accuracy



Types of Multicollinearity

The types of multicollinearity are of four types as follows:

1. High Multicollinearity: It signifies a high or strong correlation between two or more independent variables, but not a perfect one.
2. Perfect Multicollinearity: This degree of collinearity indicates an exact linear relationship between two or more independent variables.
3. Data-based Multicollinearity: The possibility of collinearity, in this case, arises out of the selected dataset.
4. Structural Multicollinearity: This issue arises when researchers have a poorly designed framework for the regression analysis.

4.4.2 Causes of Multicollinearity

Multicollinearity may be due to the following factors:

1) **The data collection method employed**
   For example, Sampling over a limited range of values taken by the regressors in the population.
2) **Constraints on the model or in the population being sampled.**
   For example in the regression of electric energy consumption on income and house sine here is a physical constraint in the population in that families with higher incomes generally have larger homes than families in lower is comes.
3) **Model Specification**
   For example adding polynomial terms to a regression model, especially when the range of the 'X'-variable is small.
4) **An over determined model**
   This happens when the model has more explanatory variables than the number of observations. This would happen in medical research, health economics where there may be a small number of patients about whom information is collected on large no of variables.

Consequence ofMulticollinearity.

The main consequences of multicollinearity are the following

1) The precision of estimation falls, so that it becomes very difficult, to separate the relative influence of various X variables This loss of precision has three aspects;
   (a) Specific estimates may have very large errors,
   (b) These errors may be highly correlated and
   (c) the sampling variances of the coefficients may be very large
2) Investigators are sometimes led to door variables in correctly from an analysis because their coefficients are not significantly different from zero but the true situations may be that a variable has no effect but Simply because the set of sample data has not enabled as to pick it up.
3) Estimates of coefficients become very sensitive to particular set of sample data and the addition of a few more observations can sometimes produce dramatic shifts in some of the coefficients.

4.4.3. Detection of Multicollinearity

Here the question arises as, How does you know that collinearity is present in any given models involving more than two explanatory variables?

The answer is as follows as:

1) **High R^2 value but few significant't' ratios** which is a classic symptom of presence of Multicollinearity
2) There exist a high zero order correlation coefficient between two regressors or High pair wise correlation among regressors, then multi collinearly o existing and a serious problem too.
3) The examination of partial correlation coefficients may suggest that the explanatory variables are intercorrelated or not. But this method was criticized by John 'O' Hagan and Brendan McCabe.
4) The presence of Multicollinearity can be detected by using Auxilliary regression to the main regression of Y on the X's.
   If the computed f exceeds the critical F; at the chosen level of significance, it is taken to mean that the particular X is collinear with other X's. If it does not exceed the critical Fi, then there is no collinear with other X's.
5) The Eigenvalues and Condition index aids to diagnose multicollinearity. Condition Index is the ratio of the square root of maximum eignvalue with respect to minimum eigen value. If the condition index is between 10 and 30, there is moderate to strong multicollinearity, and if it exceeds so there is severe multicollinearity.
6) The larger the value of variance inflation factors VIF the more collinear the variable X's. The tolerance value is closer to zero, the greater the degrees of collinearity of that variable with the other regressors. If the tolerance (TOL) is to I, then there no collinearity between X_j with other regressor

A variance inflation factor (VIF) is **a measure of the amount of multicollinearity in regression analysis**. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.

Small VIF values, VIF < 3, indicate low correlation among variables under ideal conditions. The default VIF cutoff value is 5; only variables with a VIF less than 5 will be included in the model. However, note that many sources say that a VIF of **less than 10** is acceptable.

Remedial measures for Multicollinearity

The multicollinearity can be solved by two options

i) Do nothing or
ii) Follow some rules of thumb. The rules of thumb are as follows:
   a) Using extraneous or pool information,
   b) Combining cross sectional and time series data

c) Omitting a highly collinear variable,
d) Transforming data and
e) Obtaining additional or new data

4.5 Heterocedasticity

The classical linear regression model assumes that, the variance of each disturbance term U; is same for all observations. That o the conditional variances of U; are identical symbolically $\text{Var}\left(\frac{U_i}{X_i}\right) = \sigma^2$ where var stands for variance. This is the assumption of Hemoscedasticily, or equal variance. If this assumption is violated, heteroscedasticity arises

4.5.1 Meaning

Heteroscedasticity refers to for given vales of X's the variance of each disturbance term U_i is not constant number equal to σ^2

$$\text{var}\left(\frac{U_i}{X_i}\right) \neq \sigma^2$$

4.5.2 Causes Heteroscedasticity

The causes of variances of Ui may be variable are as follows:

1) The error of learning models, as people - learn their errors of behaviour become smaller over time and is expected to decrease
2) The existence of discretionary of human behaviour is an another cause of variance of Ui may be variable.
3) The method of data collection and data processing is as a cruising factors a variance varied nature
4) The presence of outliers or an outlying e observation in relation to the observation in the sample alter the results of regression analysis and causes the heteroscedasticity.
5) The misspecification of model is also a reason for heteroscedasticity
6) Hetetoscedasticity arises due to the skewness is the distribution of one or more regressors included in the model.
7) Incorrect data transformation and incorrect functional form are sources of hetroscedasticity.

Consequence ofHetroscedasticity

If the assumptions of homoscedastic, disturbance is not fulfilled, we have the following consequences:

1) We cannot apply the formula of the variances of the coefficients to conduct tests of significance and construct confidence intervals. The tests are replicable.

2) If μ is heteroscedastic, the OLS estimate do not have the minimum variance property in the Class of unbiased estimators, that is they are inefficient in small samples and large samples.
3) The presence of heteroscedasticitythat is the variance of the "U"; "S" are not constant does not require for unbiasaedness of the least square estimates. It means that the coefficients estimates would be statistically significant even in presence of heteroscedasticity.
4) The prediction of Y for given value of X based on the estimates from the original data would be have a high variance, that is the prediction would be inefficient. Because the variance of the prediction includes the variances of 'U' and of the parameter estimates, which are not minimal due to the incidence of heteroscedasticity.

4.5.3. Detecting Hetéroscedasticity

Various tests have been suggested for establishing homoscedasticity. Here in following pages, the test which are conceptually and computationally simple one to apply are presented.

1. Park Test: Park

Suggested a functional form as $\sigma_i^2 = \sigma^2 x X_i^{\beta} e^{ti}$ (or)

taking log in both sides

$$\log \sigma^2 = \log \sigma^2 + \beta \log X; + V_i$$

where V is the Stochastic disturbance term. Since σ^2 is generally not known park suggests using U^2 as a proxy and running the following regression $\log U^2 = \log X^2 + \beta \log X_i + \vartheta$;

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

2. The Spearman rank correlation test

This is the simplest test, which may be applied for small or large samples. The Steps are as follows.

1. Regress Y on X .

$Y = \alpha + \beta X_1 + U$ and obtain the residuals, $\hat{e}$'s which are estimates of the U's .

2. Arrange the x values, and the e's in ascending or descending order and compute the rank correlation Coefficient

$$r_{e.x} = 1 - \frac{6 \sum Di^2}{n(n^2 - 1)} \text{ where}$$

DI = difference between the ranks corresponding pairs of X and e.

n = observation in the sample.

3. The result of high rank correlation

Coefficient is high suggests that the presence of heterscedasticity, otherwise if the rank correlation coefficient is low means the presence of homoscedasticity

3. Gold feld - Quant Tests

The regression model

$Y_i = \alpha + \beta X_i + U_i$ Then

Step-1 Order or rank the observations according to to the values Xi, beginning with the lowest Value 'X'

Step. 2 Omit 'c' central observations, where 'c' is specified a priori, and divide the remaining (n-c) observations into two groups each of (n-c)/2 Observations.

Step.3 Fit separate, OLS regressions to me first (n-c)/2 Observations and the last (n-c)/2 observations, and obtain the respective residual sum of squares RSS_1, and. RSS_2

Residual sum square (RSS_1) corresponding to smaller X_i values and RSS_2 is from the larger X_i values.

The degree of freedom is $\frac{(n-2)}{2} - k$ where 'k' is the number of parameters to be estimated including the intercept.

Step 4 Compute the ratio

$\lambda = \frac{RSS_2/df}{RSS_1/df}$ Where df Stands for degrees of freedom.

The $\lambda = F$, so the computed λ is greater than Critical F at the chosen level of significance Then we can say that heteroscedasticity is existing in model estimation process.

Remedial measures

The presence of heteroscedasticity does not destroy the unbiasedness and consistence properties of ous estimators, but They are no longer efficient, and not even asymptotically. Therefore remedial measures needed to solve the problem of heteroscedasticity

There are two approaches to remediation: When σ_i^2 is known and when σ_i^2 is not known.

(a) When σ_i^2 is known, then to correct the problem of heterosedasticity is by means of Weighted least squares, for the estimators thus obtained are BLUE.

(b) When σ_i^2 is not known, then use the data transformation method(1) based on the assumption of the error covariance is proportional to X_i (iii) the error variance is proportional to the Square of the mean values of Y. (iv) log transformation Which Compresses the scales in which the Variables are measured ..

To conclude the above discussion of the remedial measures which of the Transformation discussed will work depends on the nature of the problem and severity ofheteroscedasticity.

4.6 Auto correlation meaning

Auto correlation is also called 'Serial correlation'. Auto, which means self, signifies that a series is correlated with itself. Auto correlation refers to the relationship between successive error terms.

But the classical linear regression model assumption is that successive disturbance terms are drawn at random that is

$E(U_i U_j)=0$ for $i \neq j = 1,2,3$ ……n.

It implies that when observations are made over time the effect of the disturbance to occuring in one period does not carry over to another period. It means no Auto correlation.

4.6.1. Causes Auto Correlation

The causes of autocorrelation are as follows:

1) Inertia or sluggishness of economic time series leads to successive observations are likely to be interdependent.
2) Misspecification of the relationship or
3) Excluding important variables, functional form.
4) Contains errors of measurement.
5) The nature of Cobweb phenomena means that decision making at a variable depends upon its past and data messaging and data transformation.

Consequences of Auto Correlation

The U is auto correlated, and then the following Consequences will be as follows as

(a) The presence of auto correlation the OLS estimators remain unbiased, Consistent and asymptotically normally distributed, but they are no longer efficient.
(b) If U's are autocorrected, then the predictions based on OLs estimators will be inefficient
(c) The confidence intervals are unnecessarily widen and the tests of significance 'E' and 'F' and X are no longer valid and if applied are likely to give seriously misleading conclusions about the statistical significance of the estimated regression coefficients.

4.6.2. Detection and Remedies.

There are formal and informal methods of detecting the presence of autocorrelation.Among the informal methods one can simply plot the actual or standardized residuals, or plot current residuals against past residuals. In formal methods, one can use the runs test, Durbin-Watson's' test, asymptotic normality test, Berenblutt - Webb test, and Boot Breusch Godfrey (BG) test. Durbin Watson'd' test is commonly used with its limitations. So, it is better to use Breusch Godfrey test.

The Remedial measures are based on the knowledge. One has about the nature of interdependence among the disturbances that is knowledge about the structure of autocorrelation. Then the remedial tan measures can be grouped as when ρ to know and ρ is not known. The problem of autocorrelation can be removed by Markov first order autoregressive scheme, known as AR(I) Scheme. When ρ is not known. This scheme assumes that the disturbance in the current time period is linearly related to the disturbance term in the precious time period, the coefficient of autocorrelation e providing the extent of the autocorrelation of providing the extent of the interdependence.

If the value of l is known, then the problem of autocorrelation can be removed by using Durbin - Watson a Theil- Nagar Modified and Cochrane - Orcutt (C-O) iterative procedure.

4.7. Specification Error

Specification of a regression refers to formulate the regression equation. Specification of a model is the logical idea of economic theory. But due to mishandling or inadequate knowledge of economic fact and economic theories, the errors occur due to this misspecification is known as "Specification Error?

4.7.1. Meaning

In simple words, Specification error means the error that occur because of mistake in variables inclusion or exclusion or assumption of the model.

Reasons for Specification errors

Specification error is common and intractable problem in economics. In reality, any economic variable is determined by a number of factors and all of which will not be included in regression analysis. The inclusion of large number of explanatory variables will of course reduce the number of degrees of freedom in the analysis, and making the estimates of the parameters in precise.

Some of the explanatory variables may not be quantifiable and there for difficult to incorporate into numerical analysis.

Some of the variables may be or rifted by mistake because their relevance is unrecognized, or may be the lack of knowledge of researcher. Hence the error of specification arises due to:

a) Omission of a relevant explanatory variable (S)
b) Inclusion of an irrelevant explanatory variable (S)
c) Discarding of a qualitative change in one of the explanatory variables and
d) Incorrect mathematical form of the regression equation.

Types of Specification error

A regression model will have a specification error when at least one of the following problem occur in that model

1. Inclusion of irrelevant explanatory variable
2. Omission of relevant explanatory variable and
3. Incorrect functional form

4.7.2 Consequences

The inclusion of irrelevant variable (S) does not affect the relationship between other variables and the dependent variable,Because, the estimator for such a variable turns out to be zero. The estimates of inclusion of irrelevant variable in a model are unbiased and consistent. However the estimates are not efficient because of the variance are larger than they would have been in the model excluding the irrelevant variable. Further the model estimators violate the properties of 'BLUE', the concept of regression because the estimators are inefficient. If the specification error is due to qualitative change in one or more explanatory variables, then also, the estimations will be biased.

Another sort of specification error arises when the functional relationship is incorrect. The magnitude of bias will depend upon the size of coefficients. Thus the estimated parameters will be biased if we calculate the parameters without taking into account of the errors committed.

Comparison of specification error of exclusion/ omission of relevant explanatory variable and inclusion of irrelevant variable.

S.No	Category of information	Omission\ Exclusion model	Inclusion model
1	Estimation of Coefficient	Biased	unbiased.
2	Efficiency	Generally declines	Declines
3	Estimation of disturbance term	over estimate	unbiased.
4	Convention test of hypothesis and confidence region	invalid and faculty inferences	valid through erroneous

Tests of specification error

To detect equation Specification errors, there are several test are used. The prime tests are (a) examination of residuals (b) the Durbin Watson d statistic (c)Ramsey's RESET test and (d) the Lagranage multiplier test

4.8 Let us Sum up.

If the assumptions of the Classical linear regression model that the errors term or disturbance term ' U_i '

(a) (i) entering into the population regression function (PRF) are random or uncorrelated, (autocorrelation) (ii) have all the same variance σ^2 (homoscedasticity) and (b) There is no linear relationship among the explanatory variables X's (multcolltinearity) are violated causes of violation what will happen in estimation? How to identify the factors influencing for violations of assumption? and what are the remedial measures to solve the issue of violation of assumption were studied in a descriptive way not in empirical way.

4.9. Unit End Exercises

A. Multiple Choice Questions (MCQ)

1. Find out which is not the violation of assumption .
   A. Autocorrelation C. Multicollinearity
   B. Heteroscedasticity D. Dummy variable
2. Homoscedasticity means that
   A. $Var(U_i/X_i) = \sigma^2$ B. $Var(U_i) = 0$ C. $Var(U_i) = 1$ D. $Var(U_i) = \infty$
3. Which of the following pair is not correctly matched?
   A. Dicky- Fuller test – Hetroscedasticity
   B. Durbin's h test – Autocorrelation in autoregressive models
   C. F test – Overall significance of the regression model
   D. Distributed lag models – Koyck approach
4. Park Test is used for what purpose?
   A. Detecting Hetroscedasticity
   B. Solving Hetroscedasticity
   C. Detecting Multi-collinearity
   D. Solving Multi-collinearity
5. E (Ui,Uj) ≠ 0, when i≠j is termed as,
   A. Auto-Correlation B. Hetroscedasticity
   C. Multi-collinearity D. Homoscedasticity
6. The IV estimator can be used to potentially eliminate bias resulting from
   A. Multicollinearity B. serial correlation
   C. errors in variables D. heteroskedasticity
7. Variance Inflation Factor is used for…
   A. Detecting Hetroscedasticity B. Solving Hetroscedasticity
   C. Detecting Multi-collinearity D. Solving Multi-collinearity
8. If the value of Durbin-Watson's d statistic = 0, there is……
   A. No Auto-correlation B. Positive Auto-correlation
   C. Negative Auto-correlation D. None of these
9. The assumption of homoscedasticity was expressed as…….
   A. $E(U_i)^2 = \sigma_i^2$ B. $E(Ui)^2 = \sigma^2$ C. $E(Ui)^2 = 0$ D. $E(Ui) = 0$
10. Which of the following is a multi-collinearity diagnostic?
    A. Condition Index. B. Park test.
    C. Glejsertes. D. Durbin's h test.
11. Which of the following is used to detect specification errors?
    A. The Park test. B. Ramsey's RESET test.
    C. Chow test. D The Runs test.
12. What is the meaning of the term 'heteroscedasticity'?
    A. The variance of the errors is not constant.
    B. The variance of the dependent variable is not constant.

C. The errors are not linearly independent of one another.
D. The errors have non-zero mean.

13. When one or more of the regressors are linear combinations of the other regressors, it is called ----------.
A. Autocorrelation. B Heteroscedastity.
C. Multicollinearity. D. Serial correlation.

14. As a rule of thumb, a variable is said to be highly collinear if the Variance Inflation Factor (VIF) is ----------.
A. Exactly 10. B Exceeds 10. C. Less than 10. D. None of the above.

15. Specification bias or specification error means
A. Leaving out important explanatory variables
B. Including unnecessary variables
C. Choosing the wrong functional form between Y and X variables
D. All of the above

Part B. Short Answers and Essay type Questions

1. Elucidate the causes, consequences and remedies of Autocorrelation.
2. Enumerate the causes, consequences and remedies of Multicollinearity.
3. Explicate the causes and consequences of Heteroscdasticiy.
4. What are diagnostic test used for Autocorrelation, Mulitcolloinerity, and Hetroscedasticity?
5. What is specification error? Compare the effects of exclusion of relevant variable with inclusion of irrelevant variable**.**

4.10 Reference Books

1. DamodarN.Gujarati and Sangeetha, " Basic Econometrics" Special Indian Edition , Tata McGraw Hill Education Privated Limited (Sixth Print 2010) , ISBN: 978-0-07-066005-2
2. P.G.Apte," Text book of Econometrics" Tata McGraw – Hill Publishing Company Limited.
3. Dhanasekaran," Econometrics" 2nd Edition, Vrinda Publications (P) Ltd, Delhi-53 ,2011 ISBN: 978-81-8281-388-5
4. HumbertoBarreto and Frank M. Howland," Introductory Econometrics" Cambridge University Press, First South Asian Edition 2009, ISBN: 978-0-521-12358-9.
5. Dilip M. Nachane," Econometrics: Theoretical Foundations and Empirical Perspectives" Oxford University Press, Second Impression 2010 , ISBN: 978-0-19-564790-7
6. S.Shyamala,RavdeepKaur, Arul Pragasam," A text book on Econometrics: Theory and Applications" Vishal Publishing Co., Jalandhar 2017, ISBN: 81-88-646-98-9
7. S.P.Singh, Anil.K. Parashar, H.P. Singh," Econometrics and Mathematical Economics" Second Revised Editions,S.Chand and Company Ltd, New Delhi -55

8. A. Koutsoyiannis," Theory of Econometrics" Second Edition Palgrave – New York,2004,ISBN: 0-333-77822-7
9. Maddala .G.S.(1997), "Econometrics" , McGraw Hill, New York.
10. Johnston. (1997)," Econometric Methods" McGraw Hill, 4th Edition, New Delhi.

UNIT V

Dynamic & Qualitative Regression Models

5.1. Objectives

- To learn and apply the knowledge in real data set by construction of econometric modeling with dummies
- To understand the lag and its reason for introduction in analysis
- To study the ANOVA and ANCOVA
- To learn regression on qualitative independent variables and qualitative dependent variables

5.2 Introduction

Econometric researches incorporate many economic variables. Some of them are quantifiable or measurable while some variables are qualitative and hence are not measurable directly. In general, the explanatory variables in any regression analysis are assumed to be quantitative in nature. For example, the variables like temperature, distance, age etc. are quantitative in the sense that they are recorded on a well-defined scale. But in reality all variables in an economic activity may not be measureable, in such qualitative cases of variables, this unit knowledge will be used for finding the qualitative variables influence on quantitative and qualitative variable. In this unit, the readers may get the knowledge of lag variable usage for to study the implications of past period and shocks. Further reader learns about the qualitative variables usage as independent and dependent in a regression analysis as dummies.

5.3 Lag and Reasons for Lag

In economics the dependence of a variable Y (the dependent variable) on another variable(s) X (the explanatory variable) is rarely instantaneous. Very often, Y responds to X with a lapse of time. Such a lapse of time is called a *lag*. There are three main reasons for lag:

1. **Psychological reasons.** As a result of the force of habit (inertia), people do not change their consumption habits immediately following a price decrease or an income increase perhaps because the process of change may involve some immediate disutility. Thus, those who become instant millionaires by winning lotteries may not change the life styles to which they were accustomed for a long time because they may not know how to react to such a windfall gain immediately. Of course, given reasonable time, they may learn to live with their newly acquired fortune. Also, people may not know whether a change is "permanent'' or "transitory.'' Thus, my reaction to an increase in my income will depend on whether or not the increase is permanent. If it is only a nonrecurring increase and in succeeding periods my income returns to its previous level, I may save the entire increase, whereas someone else in my position might decide to "live it up.''
2. **Technological reasons.** Suppose the price of capital relative to labor declines, making substitution of capital for labor economically feasible. Of course, addition of capital takes time (the gestation period). Moreover, if the drop in price is expected to be temporary, firms may not rush to substitute capital for labor, especially if they expect that after the temporary drop the price of capital may increase beyond its previous level. Sometimes, imperfect knowledge also accounts for lags. At present the market for personal computers is glutted with all kinds of computers with varying features and prices. Moreover, since their introduction in the late 1970s, the prices of most personal computers have dropped dramatically. As a result, prospective consumers for the personal computer may hesitate to buy until they have had time to look into the features and prices of all the competing

brands. Moreover, they may hesitate to buy in the expectation of further decline in price or innovations.

3. **Institutional reasons.** These reasons also contribute to lags. For example, contractual obligations may prevent firms from switching from one source of labor or raw material to another. As another example, those who have placed funds in long-term savings accounts for fixed durations such as one year, three years, or seven years are essentially "locked in'' even though money market conditions may be such that higher yields are available else where. Similarly, employers often give their employees a choice among several health insurance plans, but once a choice is made, an employee may not switch to another plan for at least one year. Although this may be done for administrative convenience, the employee is locked in for one year

For psychological, technological, and institutional reasons, a regress and may respond to a regressor(s) with a time lag. Regression models that take into account time lags are known as **dynamic** or **lagged regression models.** There are two types of lagged models: **distributed-lag** and **autoregressive.** In the former, the current and lagged values of regressors are explanatory variables. In the latter, the lagged value(s) of the regress and appears as an explanatory variable(s).

5.4 Distributed Lag Model (DL)

In regression analysis involving time series data, if the regression model includes not only the current but also the lagged (past) values of the explanatory variables (the X's), it is called a **distributed-lag model.** Thus, $Yt = a + \beta 0Xt + \beta 1Xt-1 + \beta 2Xt-2 + ut$ represents a distributed-lag model

Autoregressive Model (AR)

If the model includes one or more lagged values of the dependent variable among its explanatory variables, it is called an **autoregressive model.**

$Yt = a + \beta Xt + \gamma\ Yt-1 + ut$

is an example of an autoregressive model. It also known as **dynamic models** since they portray the time path of the dependent variable in relation to its past value(s).

Estimation of Distributed Lag Model (DL)

A purely distributed-lag model can be estimated by OLS, but in that case there is the problem of multi collinearity since successive lagged values of a regress or tend to be correlated. As a result, some shortcut methods have been devised. These include the Koyck, the Adaptive expectations, and partial adjustment mechanisms, the first being a purely algebraic approach and the other two being based on economic principles. Aunique feature of the **Koyck, adaptive expectations,** and **partial adjustment models** is that they all are autoregressive in nature in that the lagged value(s) of the regress and appears as one of the explanatory variables.

5.5. Adhocestimation suffers from many drawbacks, such as the following:

1. There is no a priori guide as to what is the maximum length of the lag.
2. As one estimates successive lags, there are fewer degrees of freedom left, making statistical inference somewhat shaky. Economists are not usually that lucky to have a long series of data so that they can go on estimating numerous lags.
3. More importantly, in economic time series data, successive values (lags) tend to be highly correlated; hence multi collinearity rears its ugly head. Multi collinearity leads to imprecise estimation; that is, the standard errors tend to be largein relation to the estimated coefficients. As a result, based on the routinely computed *t*ratios, one may tend to declare (erroneously), that a lagged coefficient(s) is statistically insignificant.
4. The sequential search for the lag length opens the researcher to the charge of **data mining.**
5. In view of the preceding problems, the adhoc estimation procedure has very little to recommend it.

5.6 Koyck transformation

Koyck has proposed an ingenious method of estimating distributed-lag models. *Assuming that the β's are all of thesame sign*, Koyck assumes that they decline geometrically as follows

$\beta_k = \beta_0\lambda^k$ *where* $k = 0, 1, \ldots$Eq. (1)

where λ, such that $0 < \lambda < 1$, is known as the *rate of decline,* or *decay,* of the distributed lag and where $1 - \lambda$ is known as the *speed of adjustment*. Eq. (1) postulates is that each successive β coefficient is numerically less than each preceding β (this statement follows since $\lambda < 1$), implying that as one goes back into the distant past, the effect of that lag on Yt becomes progressively smaller, a quite plausible assumption. After all, current and recent past incomes are expected to affect current consumption expenditure more heavily than income in the distant past.

Equation one may be written as

$Yt = a + \beta 0Xt + \beta 0\lambda Xt-1 + \beta 0\lambda 2Xt-2 + \cdots + ut$

$Yt-1 = a + \beta 0Xt-1 + \beta 0\lambda Xt-2 + \beta 0\lambda 2Xt-3 + \cdots + ut-1$

$\lambda Yt-1 = \lambda a + \lambda\beta 0Xt-1 + \beta 0\lambda 2Xt-2 + \beta 0\lambda 3Xt-3 + \cdots + \lambda ut-1$

$Yt - \lambda Yt-1 = a(1 - \lambda) + \beta 0Xt + (ut - \lambda ut-1)$

$Yt = a(1 - \lambda) + \beta 0Xt + \lambda Yt-1 + vt$

Where $vt = (ut - \lambda ut-1)$, a **moving average of *ut* and *ut*−1.**The procedure just described is known as the **Koyck transformation.**

The following features of the Koyck transformation:

1. Any One Researcher started with a distributed-lag model but ended up with an autoregressive model because $Yt-1$ appears as one of the explanatory variables. This transformation shows how one can "convert" a distributed-lag model into an autoregressive model.

2. The appearance of $Yt-1$ is likely to create some statistical problems. $Yt-1$, like Yt, is stochastic, which means that we have a stochastic explanatory variable in the model. Recall that the classical least-squares theory is predicated on the assumption that the explanatory variables either are non stochastic or, if stochastic, are distributed independently of the stochastic disturbance term. Hence, we must find out if $Yt-1$ satisfies this assumption.
3. In the original model the disturbance term was ut, whereas in the transformed model it is $vt = (ut - \lambda ut-1)$. The statistical properties of vt depend on what is assumed about the statistical properties of ut, for, as shown later, if the original ut's are serially uncorrelated, the vt's are serially correlated. Therefore, we may have to face up to the serial correlation problem in addition to the stochastic explanatory variable $Yt-1$.
4. The presence of lagged Y violates one of the assumptions underlying the Durbin–Watson d test. Therefore, we will have to develop an alternative to test for serial correlation in the presence of lagged Y. One alternative is the **Durbin *h* test,**

Auto regressiveness poses estimation challenges; if the lagged regress and is correlated with the error term, OLS estimators of such models are not only biased but also are inconsistent. Bias and inconsistency are the case with the Koyck and the adaptive expectations models; the partial adjustment model is different in that it can be consistently estimated by OLS despite the presence of the lagged regress and.

To estimate the Koyck and adaptive expectations models consistently, the most popular method is the **method of instrumental variable.** The instrumental variable is aproxy variable for the lagged regress and but with the property that it is uncorrelated with the error term.

An alternative to the lagged regression models just discussed is the **Almon polynomial distributed-lag model,** which avoids the estimation problems associated with theautoregressive models. The major problem with the Almon approach, however, is thatone must *prespecify*both the lag length and the degree of the polynomial. There areboth formal and informal methods of resolving the choice of the lag length and thedegree of the polynomial.

5.7 Dummy Variable (qualitative variable as explanatory)

In regression analysis the dependent variable, or regressand, is frequently influenced notonly by ratio scale variables (e.g., income, output, prices, costs, height, temperature) butalso by variables that are essentially qualitative, or nominal scale, in nature, such as sex,race, color, religion, nationality, geographical region, political upheavals, and party affiliation.For example, holding all other factors constant, female workers are found to earn lessthan their male counterparts or nonwhite workers are found to earn less than whites. Thispattern may result from sex or racial discrimination, but whatever the reason, qualitativevariables such as sex and race seem to influence the regressand and clearly

should beincluded among the explanatory variables, or the regressors.Since such variables usually indicate the presence or absence of a "quality" or an attribute.

5.7.1 Meaning of Dummy Variable

A dummy variable is a variable that takes values of 0 and 1, where the values indicate the presence or absence of feature variables. Where a categorical variable has more than two categories, it can be represented by a set of dummy variables, with one variable for each category. Numeric variables can also be a dummy coded to explore nonlinear effects. Dummy variables are also known as indicator variables, design variables, contrasts, one-hot coding, and binary basis variables.

Technically, dummy variables are dichotomous, quantitative variables; they can take on any two quantitative values. As a practical matter, regression results are easier to interpret when dummy variables take on two specific values, 1 or 0. Typically, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

5.7.2 Nature of Dummy Variables

1. A dummy variable can only take on 2 values (0 or 1), we call the condition in which the dummy variable is 0 the base condition. Dummy variables are discrete variables taking a value of '0' or '1'. They are often called 'on' 'off' variables, being 'on' when they are 1
2. Dummy variables can be used either as explanatory variables or as the dependent variable. When they act as the dependent variable there are specific problems with how the regression is interpreted, however when they act as explanatory variables they can be interpreted in the same way as other variables.
3. The coefficient of the dummy variable represents the difference between being in the base condition and *not* being in the base condition.
4. The dummy variable affects the intercept of the regression model, not the slope

5.7.3.Types of Dummy Variables

1. Qualitative dummy variables: i.e. age, sex, race, health.
2. Seasonal dummy variables: depends on the nature of the data, so quarterly data requires three dummy variables etc.
3. Dummy variables that represent a change in policy:
   (a) Intercept dummy variables, that pick up a change in the intercept of the regression
   (b) Slope dummy variables, that pick up a change in the slope of the regression
   Broadly speaking this specially designed variable represents the following effects based on their types of dummy variable utilized in a regression analysis.
      (i) Temporal Effects: An investigator sometimes finds that behavioral relationship shifts from one period to another. As sales receipts of a shopkeeper might have a tendency to increase during first week of every month; Government expenditure might be expected to shoe an upward shift

during war period; consumption expenditure might also change during war period. Similarly, these might be temporary change in relation during the different seasons, periods or even during different political regimes.

(ii) Spatial Effects: Sometimes economic functions change with a change in country, economic structure or other regional differences. For example, consumption functions for U.S.A include some variables but when this consumption function is applied to the Indian population, necessary corrections should be made before hand. The reason is that behaviour pattern of American consumption will certainly be different from that of their Indian counter-parts. They would also be facing an environment different from that of Indian consumers. Thus consumption function for India may include the effects of different economic setting.

(iii) Qualitative Variable's Effects: Economic behaviour is also influenced by qualitative phenomena such as sex, occupation social status, material status etc. For example, consumption pattern of a newly married couple is bound to be different from that of an elderly couple. Similarly, the expenditure of white collared lobourers might be different from that of manual labourers. Thus these effects must be incorporated in the estimation process.

Effect of all above causes can be incorporated into our regression model by the specification of appropriate dummy variables. In practice we find several types of models containing dummy variables.

5.7.4 Caution in the Use of Dummy Variables

Although they are easy to incorporate in the regression models, one must use the dummy variables carefully. In particular, consider the following aspects

1. If a qualitative variable has m categories, introduce only $(m - 1)$ dummy variables. If anyone do not follow will fall in the problem of dummy variable trap, that is, the situation of perfect collinearity or perfect multicollinearity, if there is more than one exact relationship among the variables For each qualitative regressor, the number of dummy variables introduced must be one less than the categories of that variable.
2. The category for which no dummy variable is assigned is known as the base, benchmark, control, comparison, reference, or omitted category. And all comparisons are made in relation to the benchmark category.
3. The intercept value ($\beta 1$) represents the *mean value* of the benchmark category.
4. The coefficients attached to the dummy variables in a Equation are known as the differential intercept coefficients because they tell by how much the value of the category that receives the value of 1 differs from the intercept coefficient of the benchmark category.

5. If a qualitative variable has more than one category, the choice of the benchmark category is strictly up to the researcher. Sometimes the choice of the benchmark is dictated by the particular problem at hand.
6. if a model has several qualitative variables with several classes, introduction of dummy variables can consume a large number of degrees of freedom. Therefore, one should always weigh the number of dummy variables to be introduced against the total number of observations available for analysis.

5.8 ANOVA and ANCOVA

ANOVA, explicated by Ronald A. Fisher (1924, 1932, 1935b) for to analyses the data obtained from agricultural experiments, for to compare the means of any number of experimental groups or conditions without increasing the Type I error rate. Fisher (1932) also described ANCOVA with an approximate adjusted treatment sum of squares, before describing the exact adjusted treatment sum of squares a few years later (Fisher, 1935b, and see Cox and McCullagh, 1982, for a brief history). In early recognition of his work, the F-distribution was named after him by G.W. Snedecor (1934). ANOVA procedures culminate in an assessment of the ratio of two variances based on a pertinent F-distribution and this quickly became known as an F-test.

5.8.1 Meaning ANOVA

ANOVA expands to the Analysis of **V**ariance, is described as a statistical technique used to determine the difference in the means of two or more populations, by examining the amount of variation within the samples corresponding to the amount of variation between the samples. It bifurcates the total amount of variation in the dataset into two parts, i.e. the amount ascribed to chance and the amount ascribed to specific causes.

Dummy variables can be incorporated in regression models just as easily as quantitativevariables. As a matter of fact, a regression model may contain regressors that are all exclusivelydummy, or qualitative, in nature. Such models are called **Analysis of Variance(ANOVA) models.**ANOVA models are used to assess the statistical significance of the relationship between a quantitativeregressand and qualitative or dummy regressors. They are often used to compare the differencesin the mean values of two or more groups or categories, and are therefore more general than the *t*test, which can be used to compare the means of two groups or categories only.

5.8.2. Types of ANOVA

It is a method of analysing the factors which are hypothesised or affect the dependent variable. It can also be used to study the variations amongst different categories, within the factors, that consist of numerous possible values. It is of two types:

a. One way ANOVA: When one factor is used to investigate the difference amongst different categories, having many possible values.
b. Two way ANOVA: When two factors are investigated simultaneously to measure the interaction of the two factors influencing the values of a variable

consider the following model:

$Yi = \beta 1 + \beta 2D2i + \beta 3iD3i + ui$

whereYi = (average) salary of public school teacher in state i

$D2i$ = 1 if the state is in the Northeast or North Central

= 0 otherwise (i.e., in other regions of the country)

$D3i$ = 1 if the state is in the South

= 0 otherwise (i.e., in other regions of the country

5.8.3 Advantages and Disadvantages of ANOVA

Advantages of ANOVA

1. Whereas the Z test can only be used to compare the means of two populations, the ANOVA test can be used to compare the means of three or more populations.
2. If there are two different treatments/factors affecting the dependent variable, then we can use the two way ANOVA test to analyse the effect due to each treatment. The test will tell us whether the difference due to each of the treatments is significnt or not.
3. We can check equality of three or more populations means by repeatedly applying Z test pairwise. But this increases the Type 1 error. On the other hand, the same comparison done by the ANOVA technique has low Type 1 error. This means that ANOVAtest is a statistically powerful test.
4. The ANOVA method is used in clinical testing to check for the effectiveness of experimental medicines.
5. The calculations involved in calculating the F statistics are easy and involve elementary operations such as squaring, summing up and dividing. The decision criteria for rejecting or accepting the null hypothesis are easy to understand.

Disadvantages of ANOVA

1. It often happens that the parent populations do not follow the normal distribution. For example, the lifetimes of products generally follow the Weibull distribution. In such cases the ANOVA method cannot be used. For instance, we may not be able to use the ANOVA technique to compare the mean life of bulbs produced by three companies.
2. If there are two or more *dependent variables* then the ANOVA technique cannot be applied. The MANOVA test must be used in such cases.
3. It rarely happens that all the population variances are equal. If the assumption of homoscedasticity is violated then the use of ANOVA cannot be justified.
4. If the null hypothesis is rejected we can only conclude that some population means are unequal. The ANOVA test does not tell us anything about which of them are unequal. Some post hoc tests must be carried out in order to know about that.
5. Checking all the background assumptions such as independence, normality, homoscedasticity, etc. is in and of itself a difficult task.

6. Although the calculations involved are elementary they are still tedious to peform by hand. But ANOVA tests are usually carries out using statistical software so this is not a huge barrier.

5.8.4 Meaning of ANCOVA

ANCOVA stands for Analysis of Covariance, is an extended form of ANOVA, that eliminates the effect of one or more interval-scaled extraneous variable, from the dependent variable before carrying out research. Regression models containing an admixture of quantitative and qualitative variables are called **analysis of covariance (ANCOVA) models.** ANCOVA models are an extension of the ANOVA models in that they provide a method of statistically controlling the effects of quantitative regressors, called **covariates** or **control variables,** in a model that includes both quantitative and qualitative, or dummy, regressors.

When in a set of independent variable consist of both factor (categorical independent variable) and covariate (metric independent variable), the technique used is known as ANCOVA. The difference in dependent variables because of the covariate is taken off by an adjustment of the dependent variable's mean value within each treatment condition.

$Yi = \beta 1 + \beta 2D2i + \beta 3D3i + \beta 4Xi + ui$

whereYi = average annual salary of public school teachers in state ($)

Xi = spending on public school per pupil ($)

$D2i$ = 1, if the state is in the Northeast or North Central

= 0, otherwise

$D3i$ = 1, if the state is in the South

= 0,otherwise

5.8.5. Assumptions of ANCOVA

This technique is appropriate when the metric independent variable is linearly associated with the dependent variable and not to the other factors. It is based on certain assumptions which are:

a. There is some relationship between dependent and uncontrolled variable.
b. The relationship is linear and is identical from one group to another.
c. Various treatment groups are picked up at random from the population.
d. Groups are homogeneous in variability.

5.8.6 Advantages and Disadvantages of ANCOVA

Advantages of ANCOVA

Advantages of ANCOVA include better power, improved ability to detect and estimate interactions, and the availability of extensions to deal with measurement error in the covariates.

a. A covariate can be identified and used after the fact to save an experiment when significance is just being missed. Examples: in educational research, can use IQ or achievement test scores taken before experiment

b. ANCOVA is more precise than blocking if the correlation between the covariate and the criterion is greater than .6Remember, ANCOVA not only reduces bias, but it also improves sensitivity

Disadvantages of ANCOVA

The main disadvantage of ANCOVA is the underlying assumption of no difference across groups or treatment arms in terms of the covariate used in the analysis and the homogeneity of regression slopes

a. More assumptions to be violated with ANCOVA and effects of violations of those assumptions not always clear.
b. Skill of Computational labor needed, if doing by hand, takes laborious.
c. Blocking is more precise than ANCOVA when correlation between covariate and criterion is less than .4

5.8.7 Comparison of ANOVA and ANCOVA

a. The technique of identifying the variance among the means of multiple groups for homogeneity is known as Analysis of Variance or ANOVA. A statistical process which is used to take off the impact of one or more metric-scaled undesirable variable from dependent variable before undertaking research is known as ANCOVA.
b. While ANOVA uses both linear and non-linear model. On the contrary, ANCOVA uses only linear model.
c. ANOVA entails only categorical independent variable, i.e. factor. As against this, ANCOVA encompasses a categorical and a metric independent variable.
d. A covariate is not taken into account, in ANOVA, but considered in ANCOVA.
e. ANOVA characterises between group variations, exclusively to treatment. In contrast, ANCOVA divides between group variations to treatment and covariate.
f. ANOVA exhibits within group variations, particularly to individual differences. Unlike ANCOVA, that bifurcates within group variance in individual differences and covariate.

The above explanation of difference between ANOVA and ANCOVA can be put in tabular form for easy understanding and put it in memory for the readers based on six bases as follows:

Comparison of ANOVA and ANCOVA

BASIS FOR COMPARISON	ANOVA	ANCOVA
Meaning	ANOVA is a process of examining the difference among the means of multiple groups of data for homogeneity.	ANCOVA is techniques that remove the impact of one or more metric-scaled undesirable variable from dependent variable before undertaking research.

Uses	Both linear and non-linear model are used.	Only linear model is used.
Includes	Categorical variable.	Categorical and interval variable.
Covariate	Ignored	Considered
BG variation	Attributes Between Group (BG) variation, to treatment.	Divides Between Group (BG) variation, into treatment and covariate.
WG variation	Attributes Within Group (WG) variation, to individual differences.	Divides Within Group (WG) variation, into individual differences and covariate.

5.9 The Qualitative Response Models

Qualitative response regression models refer to models in which the response, or regressand, variable is not quantitative or an interval scale. The simplest possible qualitative response regression model is the binary model in which the regressand is of the yes/no or presence/absence type.

Linear Probability Model:

The simplest possible binary regression model is the linear probability model (LPM)in which the binary response variable is regressed on the relevant explanatory variablesby using the standard OLS methodology. Simplicity may not be a virtue here, forthe LPM suffers from several estimation problems. Even if some of the estimationproblems can be overcome, the fundamental weakness of the LPM is that it assumesthat the probability of something happening increases linearly with the level of the regressor.LPM is plagued by several problems, such as (1) non-normality of *ui*, (2) heteroscedasticity of *ui*, (3) possibility of *Yˆi*lying outside the 0–1 range, and (4) thegenerally lower *R*2 values

Limitation of LPM

1. The error term is not normally distributed; it also follows the Bernoulli distribution.
2. The variance of the error term is heteroskedastistic. The variance for the Bernoulli distribution is $p(1-p)$, where p is the probability of a success.
3. The value of the R-squared statistic is limited, given the distribution of the LPMs.
4. Possibly the most problematic aspect of the LPM is the non-fulfilment of the requirement that the estimated value of the dependent variable y lies between 0 and 1.
5. One way around the problem is to assume that all values below 0 and above 1 are actually 0 or 1 respectively
6. An alternative and much better remedy to the problem is to use an alternative technique such as the Logit or Probit models.
7. The final problem with the LPM is that it is a linear model and assumes that the probability of the dependent variable equalling 1 is linearly related to the explanatory variable.

For example if we have a model where the dependent variable takes the value of 1 if a student has extension contact and 0 otherwise, regressed on the student education level. The probability of contacting an extension employer will rise as education level rises.

LPM model example

The following model of Machine Learning (ML) was estimated, with Python Knowledgemarks(d) and Econometric Course Education marks (e) as the explanatory variables. Regression using OLS gives the following result.

$$\hat{t}_i = 3.12 + 0.6e_i - 0.12d_i$$
$$(2.10) \quad (0.06) \quad (0.04)$$
$$R^2 = 0.25, DW = 1.78$$
$$t = \begin{cases} 1 - ML \\ 0 - Not\ Ml \end{cases}$$

The coefficients are interpreted as in the usual OLS models, i.e. a 1% rise in econometric course educationmarks , gives a 0.60% increase in the probability of Machine learning technology adoption.The R-squared statistic is low, but this is probably due to the LPM approach, so we would usually ignore it. The t-statistics are interpreted in the usual way.

Logit Model

In the logit model the dependent variable is the log of the odds ratio, which is a linearfunction of the regressors. The probability function that underlies the logit model is thelogistic distribution. If the data are available in grouped form, we can use OLS toestimate the parameters of the logit model, provided we take into account explicitly theheteroscedastic nature of the error term. If the data are available at the individual, ormicro, level, nonlinear-in-the-parameter estimating procedures are called for.

Features of the Logit model

1. As P goes from 0 to 1 (i.e., as Z varies from $-\infty$ to $+\infty$), the logit$L$ goes from $-\infty$ to $+\infty$. 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 incontrast with the LPM model where the probabilities increase linearly with X.
3. Although we have included only a single X variable, or regressor, in the precedingmodel, 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 regressand equals 1 (meaning some event of interest happens) increases.If L is negative, the odds that the regressand equals 1 decrease as the value of X increases.To put it differently, the logit becomes negative and

increasingly large in magnitude as theodds ratio decreases from 1 to 0 and becomes increasingly large and positive as the oddsratio increases from 1 to infinity.

5. More formally, the interpretation of the logit model given in Eq.is as follows:

β2, the slope, measures the change in L for a unit change in X, that is, it tells how the log-oddsin favor of a categorical variable (owning a house) change as (income)explanatory variable changes by a unit, say, $1,000. Theintercept β1 is the value of the log-odds in favor of owning a house if income is zero. Likemost 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 oddsin favor of owning a house but the probability of owning a house itself, this can be done directlyfrom Eq. once the estimates of β1 and β2 are available.
7. Whereas the LPM assumes that Pi is linearly related to Xi, the logit model assumesthat the log of the odds ratio is linearly related to Xi .

5.10 Let us Sum up

From this unit the reader may learned briefly about the concepts used in dynamic model, reasons for using lag in a regression model, usage of dummies in regression and theoretical background about the ANOVA and ANCOVA , linear probability model and logit model**.** The basic knowledge obtained from this unit may help and induce the reader to go further reading in detailed way of these topics for their academic needs.

5.11. Unit End Exercises

A. Multiple Choice Questions

1. Including relevant lagged values of the dependent variable on the right hand side of a regression equation could lead to which one of the following?
   A. Biased but consistent coefficient estimate
   B. Biased and inconsistent coefficient estimate
   C. Unbiased but inconsistent coefficient estimate
   D. Unbiased and consistent but inefficient coefficient estimate
2. If in our regression model, one of the explanatory variables included is the lagged value of the dependent variable, then the model is referred to as
   A. Best fit model C. Dynamic model
   B. Autoregressive model D. First-difference form
3. Regressionn models containing a mixture of quantitative and qualitative variables are called :
   A. ANOVA models. C. ANCOVA models.
   B. Parallel regressions. D. Coincident regressions**.**
4. In Linear Probability Model, the:
   A. Regressand is dichotomous C. Regressand is ordinal variable
   B. Regressor is dichotomous D. Regressor is ordina
5. Which of the following models is used to regress on dummy dependent variable ?

A.The LPM model. B. The tobit model.
C. The logit model. D. All of the above.

6. A binary variable is often called as ---------- in econometrics.
   A. dummy variable. C. dependent variable.
   B. residual. D. power of a test.
7. The difference between the Autoregressive and Distributed lag model is in its lag of
   A. Dependent variable placing among explanatory variable
   B. Dependent variable
   C. Independent variable
   D. Dummy variable.
8. The linear probability model is
   A. the application of the multiple regression model with a continuous left-hand sidevariable and a binary variable as at least one of the regressors.
   B. an example ofprobit estimation.
   C. another word for logit estimation.
   D. the application of the linear multiple regression model to a binary dependentvariabl
9. Logistic regression is used when you want to:
   A. Predict a dichotomous variable from continuous or dichotomous variables.
   B. Predict a continuous variable from dichotomous variables.
   C. Predict any categorical variable from several other categorical variables.
   D. Predict a continuous variable from dichotomous or continuous variables.

PART B
Short answer and Essay type questions

1. What is Lag in Regression?
2. Write a note DL, AR, and MA
3. Find the difference between ANOVA and ANCOVA
4. Clarify the reasons for Lags in Econometric Models.
5. What are reasons for the introduction dummy variable and what are the cautions in use of dummies in a regression?
6. Enumerate the advantage and disadvantages of ANOVA and ANCOVA
7. What are the features of Logit Model?

5.12 Reference Books

1. DamodarN.Gujarati and Sangeetha, " Basic Econometrics" Special Indian Edition , Tata McGraw Hill Education Privated Limited (Sixth Print 2010) , ISBN: 978-0-07-066005-2
2. P.G.Apte," Text book of Econometrics" Tata McGraw – Hill Publishing Company Limited.
3. Dhanasekaran," Econometrics" 2nd Edition, Vrinda Publications (P) Ltd, Delhi-53 ,2011 ISBN: 978-81-8281-388-5

4. HumbertoBarreto and Frank M. Howland," Introductory Econometrics" Cambridge University Press, First South Asian Edition 2009, ISBN: 978-0-521-12358-9.
5. Dilip M. Nachane," Econometrics: Theoretical Foundations and Empirical Perspectives" Oxford University Press, Second Impression 2010 , ISBN: 978-0-19-564790-7
6. S.Shyamala,RavdeepKaur, Arul Pragasam," A text book on Econometrics: Theory and Applications" Vishal Publishing Co., Jalandhar 2017, ISBN: 81-88-646-98-9
7. S.P.Singh, Anil.K. Parashar, H.P. Singh," Econometrics and Mathematical Economics" Second Revised Editions,S.Chand and Company Ltd, New Delhi -55
8. A. Koutsoyiannis," Theory of Econometrics" Second Edition Palgrave – New York,2004,ISBN: 0-333-77822-7
9. Maddala .G.S.(1997), "Econometrics" , McGraw Hill, New York.
10. Johnston. (1997)," Econometric Methods" McGraw Hill, 4th Edition, New Delhi.

Additional Unit 6 (Not included in Syllabus MKU)

Introduction to Econometric Software Package Gretl

Structure

Econometrics requires skills in (licensed) software packages for testing the existing theory, creating a new theory, for evaluating any policy implications in an economy, for examining the impact on public by government funded programme accessibility, and utilization in a nation and for to study the cause and effect of a fact.

6.1 Objectives

This Unit is designed to provide students with the basic tools to work with data using the open source package gretl. In this Chapter students will learn how

- To write a script file.
- To Import native-type data sets and several other types of data sets.
- To explore your data.
- To Run basic statistical tests and to Run OLS regressions.
- To create graphs and plots.

6.2 Introduction to Gretl

GRETL is a useful tool and free software for teaching econometrics. GRETL has been written by Allin Cottrell based on ESL (Econometrics Software Library) code written by RamuRamanathan of the University of California, San Diego. It can be obtained from the World Wide Web at http://gretl.sourceforge.net/, where the source package and binary distributions running on GNU/Linux and Microsoft Windows in the form of a self-extracting executable can be downloaded.

GRETL is the first complete econometric software package released under the GNU software license. The software consists of a shared library, a command-line client program, and a graphical client program. It comes with many sample data files from Greene (2000) and Ramanathan, (2002), which are immediately accessible from the menu. It supports several least-squares based statistical estimators (including two-stage least squares and panel data methods), time series models (including the Cochrane–Orcutt procedure and VARs), and some maximum likelihood methods (logit and probit). It also has built-in commands for several econometric tests (including the Chow, Hausman, and Dickey–Fuller tests). A copy of Gretl can be downloaded from the Internet at http://www.sourceforge.net. It is approximately 7.5MB in size. An important item that can be found on the Gretl window is the option for defining a new variable. Often new variables must be created.

6.3 Features of Gretl Software

Gretl is an econometrics package, including a shared library, a command-line client program and a graphical user interface.

a. **User-friendly**: Gretl offers an intuitive user interface; it is very easy to get up and running with econometric analysis.
b. **Flexible:** You can choose your preferred point on the spectrum from interactive point-and-click to complex scripting, and can easily combine these approaches.
c. **Cross-platform**: Gretl's "home" platform is Linux but it is also available for MS Windows and Mac OS X, and should work on any Unix-like system that has the appropriate basic libraries
d. **Open source**: The full source code for Gretl is available to anyone who wants to critique it, patch it, or extend it.
e. **Sophisticated:** Gretl offers a full range of least-squares based estimators, either for single equations or for systems, including vector autoregressions and vector error correction models. Several specific maximum likelihood estimators (e.g. probit, ARIMA, GARCH) are also provided natively; more advanced estimation methods can be implemented by the user via generic maximum likelihood or nonlinear GMM.
f. **Extensible**: Users can enhance gretl by writing their own functions and procedures in gretl's scripting language, which includes a wide range of matrix functions.
g. **Accurate**: Gretl has been thoroughly tested on several benchmarks, among which the NIST reference datasets.
h. **Internet ready**: Gretl can fetch materials such databases, collections of textbook data files and add-on packages over the internet.
i. **International:**Gretl will produce its output in English, French, Italian, Spanish, Polish, Portuguese, German, Basque, Turkish, Russian, Albanian or Greek depending on your computer's native language setting.

Check Your Progress:

a. What is Gretl?
b. Is Gretl Paid software?
c. Who wrote the code to Gretl?
d. Is Gretl able to open in MS Window?

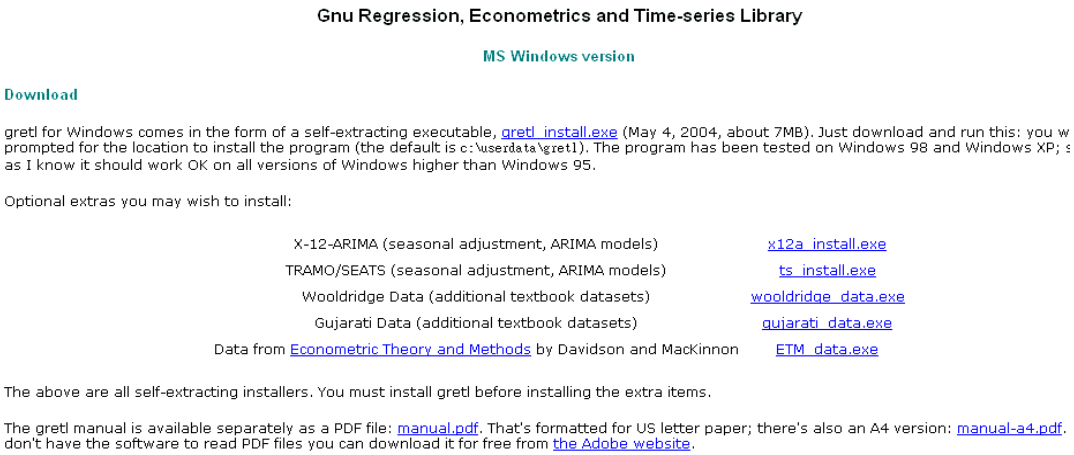
gre

6.4 Installation of Gretl Software

1. DownloadingGretl from the Internet for Free

GRETL econometric software can be downloaded at the following site:
http://gretl.sourceforge.net/

1. On the left hand side of this web site, double left-click on **"Gretl for Windows"** ,You will be directed to another web page that looks like the following:



Gnu Regression, Econometrics and Time-series Library

MS Windows version

Download

gretl for Windows comes in the form of a self-extracting executable, gretl_install.exe (May 4, 2004, about 7MB). Just download and run this: you w prompted for the location to install the program (the default is c:\userdata\gretl). The program has been tested on Windows 98 and Windows XP; s as I know it should work OK on all versions of Windows higher than Windows 95.

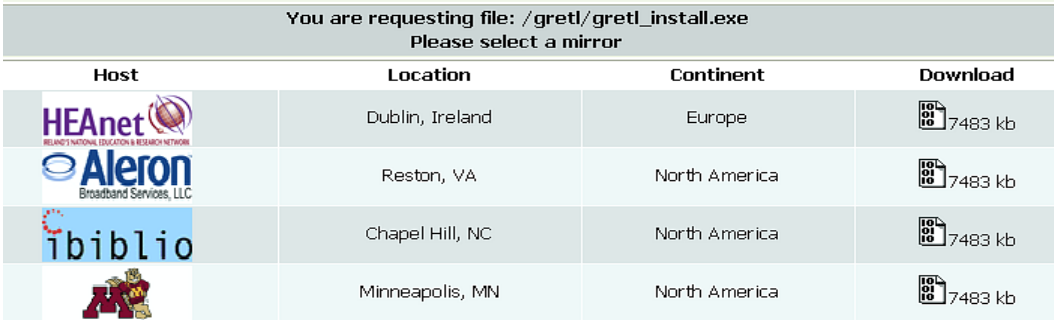
Optional extras you may wish to install:

X-12-ARIMA (seasonal adjustment, ARIMA models)	x12a_install.exe
TRAMO/SEATS (seasonal adjustment, ARIMA models)	ts_install.exe
Wooldridge Data (additional textbook datasets)	wooldridge_data.exe
Gujarati Data (additional textbook datasets)	gujarati_data.exe
Data from Econometric Theory and Methods by Davidson and MacKinnon	ETM_data.exe

The above are all self-extracting installers. You must install gretl before installing the extra items.

The gretl manual is available separately as a PDF file: manual.pdf. That's formatted for US letter paper; there's also an A4 version: manual-a4.pdf. don't have the software to read PDF files you can download it for free from the Adobe website.

2. Double left-click on **"gretl_install.exe"**which is about 7.5 MB You will be again directed to another web page like the following:



You are requesting file: /gretl/gretl_install.exe
Please select a mirror

Host	Location	Continent	Download
HEAnet	Dublin, Ireland	Europe	7483 kb
Aleron Broadband Services, LLC	Reston, VA	North America	7483 kb
ibiblio	Chapel Hill, NC	North America	7483 kb
	Minneapolis, MN	North America	7483 kb

3. Choose a **"preferred mirror"**– for example, **"unc"**as shown above.
4. Right-click on **"download 7483 kb"** say from "Chapel Hill, NC"
5. Save the file to your desktop
6. After downloading this file, open it and follow the installation wizard.
7. Gretl will create a shortcut on your desktop that looks like the following:

6.5 Creating Data Sets and Reading them into Gretl

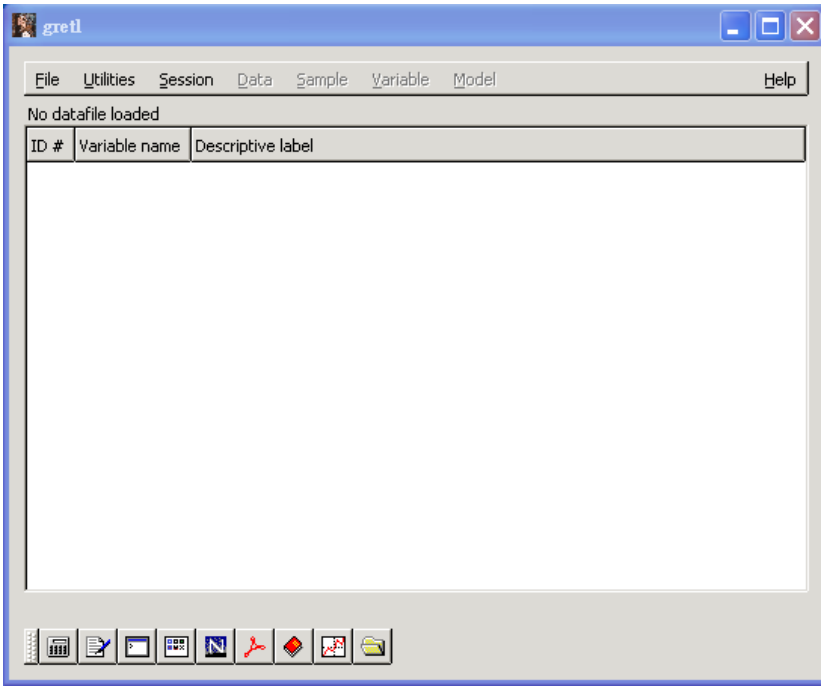
After you have successfully downloaded Gretl you will find the package was saved in the default target c:\userdata\gretl and a small icon was probably placed on your desktop. The icon looks like

You can double left-click on the gretl icon and it will start the software package. A main gretl window will open and looks like the window on the following page. The options which run along the top of the window are

File Utilities Session Data Sample Variable Model and Help

The blank part of the window will be filled when you open your **data file** (to be explained shortly). It will show a list of the variables in your data set. You can then choose the type of analysis that you wish to perform. You can run Gretl by simply clicking on options or you can write a gretl program and run it either interactively or in batch modes.

Unless you wish to explore the software on your own, the first step in a Gretl session is the creation of a data set which Gretl can understand. The simplest way this can be done is to use Notepad (or Wordpad) and the Gretl Command editor. You don't need Excel to create this kind of data set. Let's take an example.

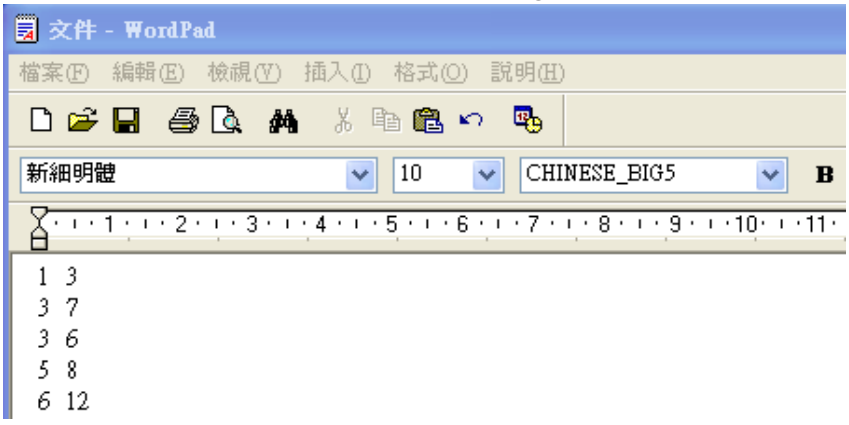


Data Set Type I: Let's suppose that you have 4 observations on two variables X1 and X2, which you can write as

Date	X1	X2
1970 Q1	1	3
1970 Q2	3	7
1970 Q3	3	6
1970 Q4	5	8
1971 Q1	6	12

You may have simply keyed these in yourself, or you may have found the data on the Internet and used a cut and paste function to create a file. Let's suppose that you use Notepad (or Wordpad) to create a data file called **mydata.txt** and which you have saved on your desktop.

The Notepad (or Wordpad) file looks like the following:



Note that this quarterly data does not have any names, like X1 or X2, in the Notepad (or Wordpad) file. It also does not have any dates, like 1970 Q3 or 1971 Q1.
Gretl will not read this data file as it is. You must do two things.

(1) You save the file again (using Notepad (or Wordpad)) as **mydat.gdt**
You should save the file to c:\userdata\gretl\user

(2) Next, you start Gretl (by clicking on the icon) and select **File** then choose
New Command File and then choose **Regular Script**

(3) A new Gretl window will open and you proceed to create a Data Header File

(4) You type into the window a description of the data and save it as **mydat.hdr .**You should save the file to c:\userdata\gretl\user

There are several things to note.

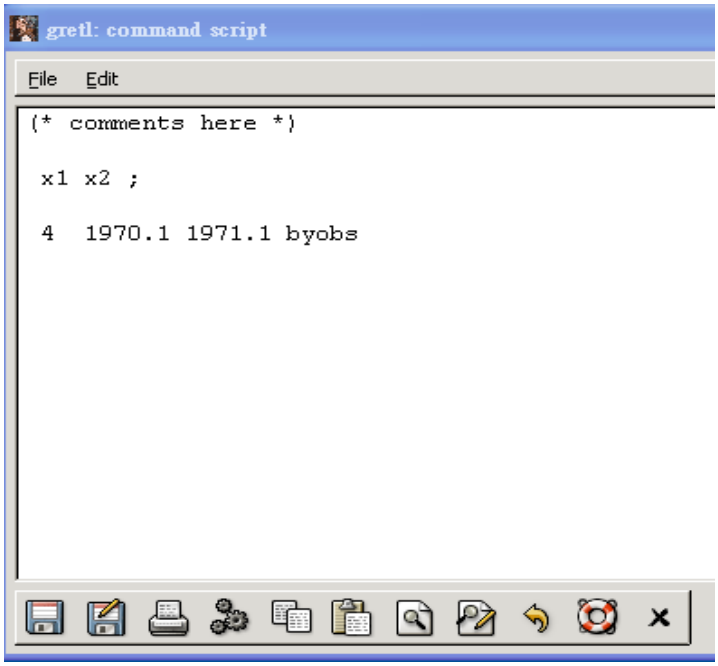
First, the data file **mydat.gdt** and the header file **mydat.hdr** BOTH use the base name "mydat"

Second, the data file must use the file suffix ___**.gdt** and the header file must use the file suffix ___**.hdr**

Third, you should save the files in the **c:\userdata\gretl\user** location.

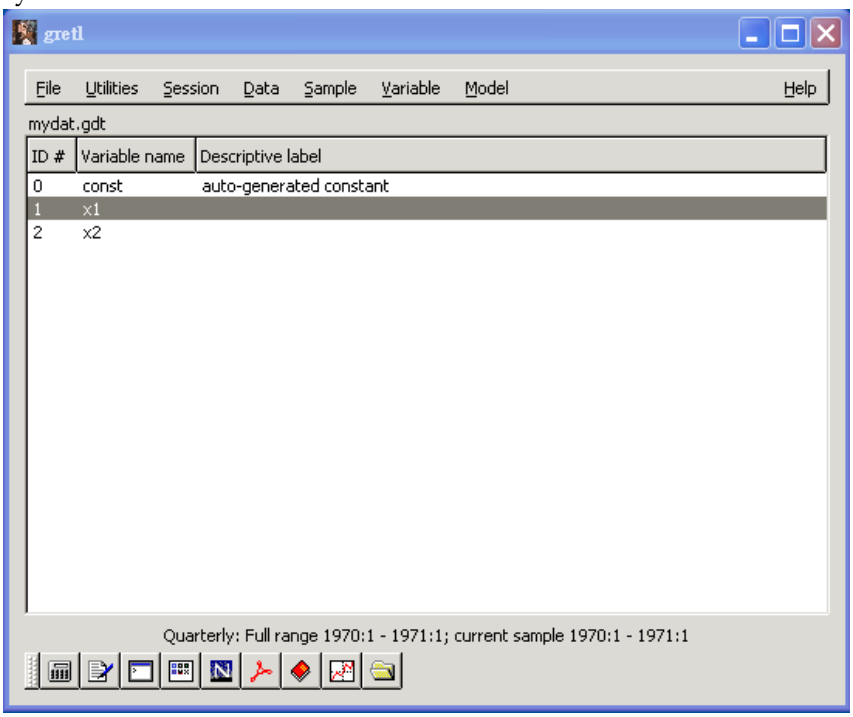
Fourth, the data is arranged in columns by observation within the data file mydat.gdt Fifth, the header file has comment lines (**(*** _____**)**), a list of names for the variables (**x1 and x2** ending in **;**) and a description of the time series nature of the data (**4**=quarterly, **1970.11971.1 byobs**).

Sixth, the variable names are case sensitive, so X1 is different from x1.



After you have created and saved both files mydat.gdt and mydat.hdr to the folder c:\userdata\gretl\user you can start Gretl and begin your analysis.

When Gretl starts (after you click on the Gretl icon), the main Gretl window will open and you can read your data into the window. To do this, you choose **FILE** and then **OPEN DATA**. You, then choose **USER FILE** and proceed to double left-click on the file **mydat** shown. Gretl will then read in your data into the main Gretl window and you will see the window shown on the following page. Note that both x1 and x2 have been read into Gretl. A constant has beenautomatically generated also. We can now carry out many types of statistical analyses on x1 and x2.



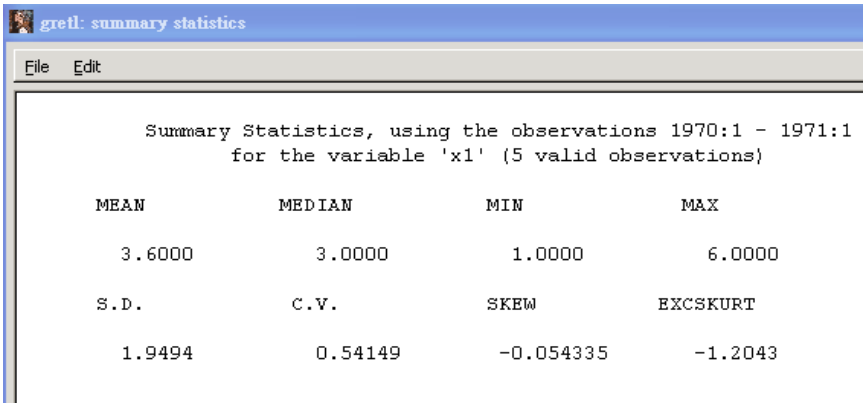
Note, also that at the bottom the data frequency and sample range is shown: "*Quarterly: Full range 1970:1 – 1971:1; current sample 1970:1 – 1971:1"* Also, at the bottom there are several convenient shortcuts

These are, respectively: (1) calculator, (2) editor (may not work), (3) interactive Gretl console, (4) icon view (must have Gretl sessions saved first), (5) Gretl website link, (6) Gretl Manual (in ___.pdf form, must have free Acrobat Reader from Adobe.com), (7) Gretl Help, (8) X-Y Graphics, (9) Open Data. There are other ways to read data into Gretl, but for now, this should be enough since it can be done on virtually any computer.

6.6 Simple Descriptive Statistics in Gretl

If you have successfully downloaded Gretl and have created the example data set above, then you can proceed to undertake simple statistical analyses of the data.

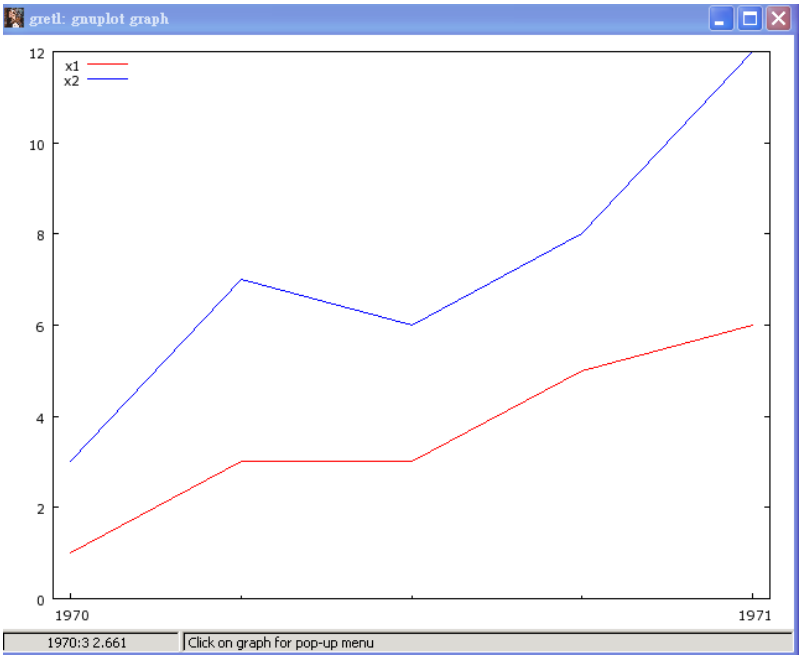
Finding the sample mean of the variable x1: Choose **Data**, **Summary Statistics**, **Selected Variables** and get the following



Note that this command gives a number of different summary statistics – not just the sample mean only. **S.D.** = standard deviation, **C.V.** = coefficient of variation = (S.D./Mean), **SKEW** = Measure of skewness, **EXCSKURT** = measure of excess kurtosis.

We can also make a time series graph for X and Y. Choose **DATA, GRAPH SPECIFIED VARS, TIME SERIES PLOT…**

Note that x1 and x2 are quarterly variables defined over the sample range 1970:1 – 1971:1. You can copy and paste this graph to a word document. Click on the graph and follow the options given. This is known as a time series graph. Gretl can also produce X-Y graphs. Just choose **DATA, GRAPH SPECIFIED VARS, X-Y SCATTER…**



6.7 Let us sum up

This unit made to introduce the free software, its features, and how to install for solving the economic problems. After installation of software in desktop or laptop, creation of data file and how to get descriptive statistics and diagrams are explained with screen.

6.8 Unit -End Exercises

1. What is Gretl? And who had written the code for it?
2. Enumerate the features of Gretl.
3. Explain the steps involved in installing Gretl in your desktop or Laptop?
4. Discuss the way of creating data file and drawing a trend line.

6.9 Answer to Check Your Progress

a. Gretl is an econometrics package, free software including a shared library, a command-line client program and a graphical user interface.
b. No. Gretl is free Software.
c. Dr.RamuRamanathan has written the code to Gretl Software.
d. Yes. Gretl is able to open in MS Window.

6.10 Suggested Readings

Gretl Used guide , and www.google.com

GLOSSARY

1. **A Cumulative Distribution Function**: A mathematical function which allows us to calculate the probability that a random variable will take on a value equal to or less than a specified value.
2. **Accounting Identity**: A relationship which holds exactly as a result of accounting conventions
3. **Adjusted R Square**: A measure of goodness of fit which is adjusted for the loss of degrees of freedom,.i.e. incorporates a penalty for using too many independent variables.
4. **Adaptive Expectation** : A model of expectation formation in which expectations are revised depending upon the discrepancy between current experience and past expectation
5. **A Just Identified Equation**: Data and prior information are just enough to uniquely identify the equation. Indirect least squares feasible.
6. **An Elementary Outcome**: One of the many possible results of a random experiment.
7. **An Event**: A happening of interest in the context of a random experiment. An event is said to have occurred when any one outcome from a specified subsets of outcomes occurs
8. **An Over-identified equation**: Data and prior restrictions apparently contain redundant information. indirect least squares leads to multiple solutions.
9. **An Unidentified Equation:**A equation from the system which cannot be distinguished form similar looking equation based on data and prior information. indirect least squares not feasible
10. **Arithmetic Lags**: Lag coefficient exhibit a linear pattern of increase or decline.
11. **An ARMA Process**: A combination of auto-regressive and moving average processes
12. **Approximate Multicollinearity**: Existence of an approximate linear relationship among the independent variables.
13. **Asymptotic Behaviour** : Behaviour of an estimator and its distribution as sample size increases without limit
14. **Asymptotic property:** A property of a statistic that applies as the sample size grows large (specifically, as it tends to infinity).
15. **Asymptotic Unbiasedness** : Bias of the estimator tends to zero as sample size increases .
16. **Attrition bias:** Bias caused by unit non-response in panel data. This occurs when the individuals who drop out of a panel study are systematically different from those who remain in a panel study.

17. **Autoregressive Process**: A stochastic process in which the current value of the disturbance is a function of past values with a random variables super imposed
18. **Average effect:** A measure of the effect of a binary explanatory variable, x, on the outcome of interest; based on comparing the outcome when x equals 1 with the outcome when X equals 0.
19. **Average treatment effect (ATE):** a measure commonly used in the policy evaluation literature that gives the expected difference in outcomes between those who receive a treatment and those who do not, across the whole study population. Related to the average treatment effect on the treated (ATET) which is the expected difference for those who would opt for treatment.
20. **Backward Elimination**: A computational routine in which one variable is dropped at a time starting from a model which includes all the independent variables.
21. **Behavioral Equation**: An algebraic relationship based some assumption about the behaviour of economic agents.
22. **Best Estimator** : within a given class of estimators the one with the minimum variance.
23. **Best Linear Unbiased Forecast**: Within the class of linear unbiased forecasts the one with minimum error variance.
24. **Bias** : The difference between the expected value of an estimator and the true value of the parameter being estimated. A measure of how well the estimator performs on average
25. **Binary variable:** A variable that takes only two values, usually coded as zero and one.
26. **Bivariate probit model:** A model that combines two binary probit models to deal with a system of two binary dependent variables.
27. **Box-Cox Transformation**: A generalised functional form which has as imiting cases several forms used often in practise. e.g. bilinear, double log, semi-log, etc.
28. **Conditional logit:** A model for unordered multinomial outcomes in which the regressors vary across the alternatives (see **mixed logit**and **multinomial logit**).
29. **Conditional Probability**: Probability of a event given that some other specified event has already occurred.
30. **Consistency**: The estimator approaches the true parameter as sample size increases.
31. **Consistent estimate:** An estimate that converges on the true parameter value as the sample size increases (towards infinity).
32. **Contemporaneous Covariance**:Across equations covariance between the disturbances referring to its same time period.
33. **Continuous variable:** A variable that can take any take the value of any real number with in an interval.
34. **Cox proportional hazard model:** A semi parametric model for duration analysis.
35. **Cross-section data:** Survey data in which each respondent is observed only once, giving a "snapshot" view of the population at a point in time.

36. **Deterministic Model**: A collection or set of exact structures, i.e. a set of autonomous relationships with parameter values unspecified.
37. **Diagnostic checking**: A set of graphical and formal testing procedures applied to OLS residuals to detect violations of basic assumptions, outliers etc.
38. **Disturbances Related Equations**: A set of equations in which disturbances in different equations are correlated.
39. **Dummy variable:** Another label for binary variables that take the value zero or one. A variable used to represent qualitative attribution and or structural breaks in regression
40. **Durbin-Watson Statistic**: A test statistic to detect the presence of first order auto-correlation. using residual from a preliminary OLS regression
41. **Distributed Lags**: Effect of a change in independent variables is spread over current and several future periods.
42. **Efficiency:** Variance of an estimator relative to that of another estimator from within a specified class of estimators. A measure of precision of an estimator.
43. **Embedding**: A procedure in which a more general specification is formed such that the two non-nested hypotheses are special cases of the general model
44. **Endogenous Variables**: Those variables the values of which are determined within the system. Otherwise, Variables the values of which are determined by or within the model.
45. **Error components model:** A regression model for panel data.
46. **Error Sum of Squares(ESS):** The sum of squares of discrepancies between observed and calculated values of the dependent variable.
47. **Estimate** : Particular numerical value of the estimator for the particular sample at hand.
48. **Estimator** : A rule for combining the sample observations to arrive at the " best guess" as to the true value of a unknown parameter.
49. **Error of Type 1**: Rejecting a valid hypothesis.
50. **Error of Type 2**: Accepting an invalid hypothesis.
51. **Extrapolation**: A procedure to estimate values outside the observed sample.
52. **Expectation of a Random Variable** : (also called mean) A measure of central tendency or "on average " behaviour of a random variable defined with its density function.
53. **Exogenous Variables**: Variables, the value of which have to be supplied from outside the model.
54. **Exact Multicollinearity**: Existence of an exact linear relation among the independent variables.
55. **Ex-Ante Forecast**: Forecasting the future values before they are actually realised.
56. **Excess zeros:** A feature of count data, when the number of zeroes observed exceeds the number that would be expected from the Poisson model.

57. **Ex-Post Forecasts**: Utilising part of the sample data to forecast values for which actual values are available.
58. **Extraneous Estimators**: Estimators of parameters in a model obtained from a different body of data in possibly a different context
59. **Exogenous Variables**: Variables which appear in the system but are determined outside the system.
60. **Exogeneity:** In the context of regression analysis, the assumption that the regressors, x, are independent of the error term.
61. **FIML:** Full-information maximum likelihood (FIML) estimates multiple equation models using the joint distribution for the equations rather than estimating each equation separately.
62. **Final Form of the Equation System**: Each endogenous variable is expressed exclusively in terms of current and lagged exogenous variables. Lagged endogenous variables eliminated.
63. **First Order Regression**: An estimation procedure for missing observations in which they are estimated from some auxiliary regressions
64. **Fixed effects:** The fixed effects specification treats the individual effects in panel data models as parameters to be estimated. This is appropriate when inferences are to be confined to the effects in the sample only, and the effects themselves are of substantive interest With individual level survey data fixed effects are best interpreted as random individual effects that are correlated with the explanatory variables. This contrasts with randome effects that are assumed to be independent of the regressors (see **random effects**).
65. **Forward Selection**:A computational routine in which one variable is added at a time starting from a model with one independent variables.
66. **Gamma distribution:** Probability distribution often used to model individual heterogeneity, especially in count data regression and duration analysis.
67. **Gauss Mark of Theorem**; The result that for the standard linear model with scalar covariance matrix of disturbances the ordinary least squares estimators are BLUE
68. **Geometric Lags**: Lag coefficients are generated from a geometric distribution and exhibit exponential decline..
69. **Gibbs sampling:** a method for drawing samples from a distribution that is used in MCMC algorithms.
70. **GMM:**Many of the estimators discussed in this book fall within the unifying framework of generalised method of moments (GMM) estimation. This replaces population moment conditions (e.g. based on expected values) with their sample analogues (e.g. based on sample means).
71. **Generalized least squares:** A generalization of ordinary least squares which relaxes the assumption that the error terms are independently and identically distributed acrossobservations.

72. **Hausman test:** Tests whether there is a significant difference between two sets of coefficients: one set that are efficient under the null but inconsistent under the alternative and another set that are inefficient under the null but still consistent under the alternative. Commonly used to test the IIA assumpition in multinomial choice models and as a test of exogeneity (comparing OLS and IV extimates).
73. **Hazard function:** Defined as the ratio of the density function to the survivor function for a random variable. The hazard function plays a key role in duration analysis where it is interpreted as the probability of failing now given survival up to now.
74. **Heckit model:** A two-step estimator designed to deal with the sample selection problem.
75. **Heteroskedasticity:** When the variance of the error term is not constant across observations.
76. **Heteroskedastic Linear Model:** A regression model in which the disturbance variance can change from observation to observation.
77. **Homoskedasticity**: The property of constant variance of random disturbance,when the variance of the error term is constant across observations.
78. **Identification:** The process of distinguishing a particular structure from a set of competing structures using data and prior information.
79. **Indirect Least Squares**: Estimating reduced form parameters by OLS and solving for structural parameters from these.
80. **Independence of Explanatory Variables and Random Disturbances**: the assumption that the explanatory variables, if they are to be treated as stochastic, are statically independent of the random disturbance.
81. **Influential Observation**: A measure of influence of an influential observation.
82. **Interval Forecast**: Analogous to interval estimation. providing an interval which will bracket the true value with stated probability.
83. **Interaction:** Joint effect of two attributes. incorporated by including products of dummy variables
84. **Interpolation**: A procedure to estimate a missing value lying between two known values.
85. **International Contribution**: Increases in the R^2 as a result of adding a variable to the model.
86. **Instrumental Variable:** A variable which is uncorrelated with the disturbance but highly correlated with the explanatory variable which it acts as the instrument
87. **Intersection of Two Events**: Joint occurrence of two events, For more than two events, the definition is same.
88. **Instability of Coefficients**: Extreme sensitivity of coefficient magnitudes and signs to small perturbations of data and/or addition of variables. a consequences of multicollinearity.

89. **Instrumental variables:** A method of estimation for models with endogenous regressors – regressors that are correlated with the error term. It relies on variables (or "instruments") that are good predictors of an endogenous regressor, but are not independently related to the dependent variable. These may be used to purge the bias caused by endogeneity.
90. **Interval regression:** A variant on the ordered probit model that can be used when the threshold values are known.
91. **Interval Estimation**: Constructing an interval which will bracket the time parameter value with a specified probability
92. **Inverse Mills ratio (IMR):** The label given to the hazard rate (ratio of density to survival functions) for a probit model. The IMR is used in the Heckit correction for sample selection bias.
93. **Inverse probability weights**: Used to re-weight sample data to make it representative of the underlying population. IPWs give more weight to those observations that are under-represented in the sample.
94. **Irrelevant Variables**: A specification error. Inclusion of unnecessary variables leads to inefficient estimates.
95. **Item non-response:** When a respondent does not provide data for a particular variable in a survey.
96. **Joint Probability**: Probability of joint, i.e. simultaneous occurrence of two or more events.
97. **Joint Density and Distribution Function** : Function specifying the joint probabilistic behaviour of one or more random variables conditional on specified events
98. **Kaplan-Meier:** A nonparametric estimator for survival curves and hazard functions.
99. **Latent Root Regression**: Similar to principal components regression but retain the predictive multicollinearities
100. **Left truncation:** A phenomenon that arises with duration data that has been sampled after the original start of the process. Left truncation occurs when some observations may have already failed before the data are collected and are therefore missing from the data.
101. **Likelihood of a Sample:** The probability that a given sample would have arisen from a particular population
102. **Limited Information Maximum Likelihood (FIML)**: A joint ML procedure for the whole system.
103. **Linear Estimator**: An estimator that is a linear function of the sample values.
104. **Linear probability model:** A model for binary dependent variables based on the linear regression model.
105. **Non-Linear Model:** Models which are linear in parameters, those which are not, are non -linear models.

106. **Logistic distribution:** A continuous probability distribution that is the foundation for the logit model of binary choice.
107. **Logit:** A model for binary dependent variables based on the logistic distribution.
108. **Maintained Hypothesis**: A set of assumptions about the phenomenon being investigated which are accepted on faith.
109. **Marginal effect:** A measure of the effect of a continuous explanatory variable, x, on the outcome of interest; based on the derivative of the outcome with respect to x.
110. **Marginal Density and Distribution Functions** : Given joint distribution these functions describe the probabilistic behaviour of one of the random variables for all possible values of the other random variables.
111. **Maximum likelihood estimation:** A method of estimation that specifies the joint probability of the observed set of data and finds the parameter values that maximize it (i.e. that are most likely). Otherwise, An approach to estimation of parameters which chooses those values for the parameters which would maximise the likelihood of the sample at hand.
112. **Mean Square Error**: Expected value of the square of the discrepancy between an estimator and the true value of the parameter.
113. **Mean Square Error of Prediction**: Mean square error of the forecast as an estimator of the actual value.
114. **MCMC:** A Bayesian method used to form a sample from the posterior density by constructing a Markov Chain in which each value is drawn conditionally on the previous iteration.
115. **Metropolis-Hastings algorithm:** A sampling method used in MCMC techniques when Gibbs sampling is not possible.
116. **Mixed logit:** A model for unordered multinomial outcomes in which the regressors can vary across individuals and across the choices. The label is also applied to the more general random parameters logit model. (See **conditional logit**and **multinomial logit**).
117. **Moving Average Process**: A stochastic process in which the current values of the disturbance term is a weighted sum of current and past values of a random variable.
118. **Multicollinearity:** Existence of linear relationship among the explanatory variables.
119. **Multinomial logit:** A model for unordered multinomial outcomes in which the regressors vary across individuals (see **mixed logit**and **conditional logit**).
120. **Negbin:** An extension of the Poisson regression model for count data.
121. **Nelson-Aalen:** A nonparametric estimator for cumulative hazard functions.
122. **Nested Hypotheses:** A pair of hypotheses in which one is a special case of the other obtained by imposing suitable restrictions on the parameters
123. **Non-Nested Hypotheses**: A pair of hypotheses in which neither can be derived as a special case of the other. also called separate families of hypotheses

124. **Normal distribution:** A continuous probabilty distribution that has a typical "bellshape". Used as the foundation for classical regression and analysis and many other models such as the probit model and the Heckit model.
125. **Null Hypothesis**: An assertion, to be tested, that the true parameter values are such and such.
126. **Omitted Variables**: A specification error. Exclusion of relevant variables leads to biased estimates
127. **Order Conditions for Identifiability:** A necessary condition stated in terms of number of restrictions imposed on the structural coefficients
128. **Ordered probit:** A model for ordered multinomial outcomes.
129. **Ordinary least squares (OLS):** The standard method for fitting the classiical linear regression model. It is based on finding the parameter values that minimize the sum of squared errors.
130. **Outlier:** An observation (data point) which appears to depart substantially from the fitted model.Otherwise, An observation which departs substantially from the rest of the data.
131. **Over-dispersion:** When observed count data are more spread out than would be expected from a Poisson model.
132. **Panel data:** Survey data in which each respondent is observed repeatedly over time.
133. **Parameters:** Unknown, unobservable constants which link up variables into a relationship.
134. **Partial Correlation**: A measure of linear association between two variables after both have been adjusted for their common dependent on a given set of other variables.
135. **Partial effect:** Used to measure the impact of a change in a regressor on the probability of the outcome of interest. Relevant for nonlinear models, such as binary choice models, where the partial effect is not simply the regression coefficient.
136. **Partial Adjustment**: A model of dynamic adjustment in which the current action is dependent upon the gap between a target and reality.
137. **Predetermined Variables**: Current exogenous, Lagged exogenous and lagged endogenous variables.
138. **Predictive Multicollinearity**: An approximate linear relationship which involves some independent variables and the dependent variable. such linear combination have good predictive power.
139. **Principal Components Regression**: An approach to estimation which uses a few linear combinations of the original variables known as principal components.
140. **Probability** : A quantitative measure of uncertainty associated with the occurrence of an event in the context of a particular experiment.

141. **Probability Density Function**: A mathematical function which gives the probability that a random variable will take on a specified value or (loosely speaking) a value within a small neighbourhood of the specified value.
142. **Point estimate:** A single number used to estimate an unknown parameter (the "best guess"). As opposed to an interval estimate, which presents a range of values.
143. **Point Estimation**: Offering a single estimate as the best guess for a parameter
144. **Point Forecast**: Analogous to point estimation. giving a single value as the best guess for the future value of the variable being forecast.
145. **Poisson regression:** A model for count data.
146. **Polynomial Lags**: Lag coefficients are generated from a polynomial of a specified degree.
147. **Pooled Data**: Data from a time series of cross sections.
148. **Probit:** A model for binary dependent variables based on the standard normal distribution.
149. **Propensity score:** The probability of participating (in a treatment) conditional on a set of regressors, $p(y=1|x)$. The propensity score is used in matching and sample selection estimators.
150. **Power of a Test:** A measure of the ability of a test to distinguish between alternative hypotheses
151. **Qualitative effect:** The sign of the effect of one variable on another.
152. **Quantitative effect:** The magnitude of the effect of one variable on another.
153. **Random disturbances**: A random variable which represents the discrepancy between an exact structure and reality
154. **Random effects:** The random effects specification treats the individual effects in panel data models as random draws. If individual effects are not of intrinsic importance in them, and are assumed to be random draws from a population of individuals, and if inferences concerning population effects and their characteristics are sought, then a random specification is suitable
155. **Random effects probit:** A model for binary dependent variables in panel data.
156. **Random Experiment**: A sequence of action carried out under specified conditions.
157. **Random Variable** : In the context of a random experiment, it is the variable which take on (real) numerical values , according to a well-defined rule based on occurrence of different events
158. **Rank Conditions for Identifiability**: A necessary and sufficient condition for an equation to be identified started in terms of rank of a sub matrix of reduced form or structural form coefficient.
159. **Raw Moments R^2 (R^2_m):** A measure of goodness of fit for models without intercept. variation in the dependent variable is measured around zero instead of its mean.
160. **Reduced form of a System**: A form in which each endogenous variables is expressed in terms of predetermined variables.

161. **Reduced Form**: A set of relationship, derived from the structural equations, in which each endogenous variable is expressed a function of all exogenous variables.
162. **RESET:** A general test for misspecification of the functional form of a regression model.
163. **Retransformation problem:** Highlights the need to use an appropriate transformation back to the y-scale when regression models are run on transformed data such as log(y).
164. **Ridge Regression**: A procedure which attempts to reduce the influence of Multicollinearities through the introduction of a biasing constant.
165. **Ridge Trace**: A graphical procedure used for choosing the value of the biasing constant in ridge regression
166. **Right censoring:** Occurs when values in the right hand tail of a distribution are cut-off at some threshold and only the threshold value is known. This often arises in duration analysis where some spells are incomplete at the time the data are collected.
167. **Risk function**: A decision -Theoretic concept. The expected cost of using an estimator to estimate a parameter. cost arises from the wrong decisions ,i.e. using an estimate different from true value
168. **Sample:** A finite collection of values of the random variable actually observed.
169. **Sampling Distribution**: Probability distribution of an estimator.
170. **Sample selection bias:** The bias created when non-responders are systematically different from responders.
171. **Sample Space**: The collection of all the possible elementary outcomes.
172. **Scalar- Covariance Matrix**: When the covariance matrix of the random disturbances is diagonal with identical diagonal elements. Consequence of homoskedasticity and serial independence.
173. **Serial Correlation**: Successive disturbance terms in the regression mode are correlated.
174. **Serial Independence**: The property of mutual stochastic independence of random disturbances.
175. **Semi parametric:** A method that mixes parametric assumptions (e.g. that the relationship between y and X is linear) and nonparametric assumptions (e.g. that the distribution of the error term is unknown).
176. **Singular- Covariance Matrix:** Contemporaneous disturbances are linearly dependent resulting in a singular covariance matrix of disturbances
177. **Specification Error Test**: A test designed to detect departure from one or more assumptions behind a proposed model.
178. **Splicing:** A procedure to obtain a consistent index series from two series with different bases but at least one point of overlap.
179. **Standardisation**: Re-expressing variables with a change of origin and units of measurement

180. **Standard Deviation**: The positive square root of the variance.
181. **Standard Multiple Correlation(R):** A measure of goodness of fit defined as the fraction of the variation in the dependent variable around its mean explained by the model.
182. **Stepwise Regression**: A procedure which allows addition and deletion of independent variables one at a time. combination of the above two procedures.
183. **Stochastic Independence**: tea event are said to be stochastically independent when the knowledge that one of them has occurred does not affect the probability of other
184. **Stochastic Model**: A deterministic model with random disturbances added on.
185. **Structure (Exact):** A collection of autonomous relationship (i.e.of the above three types) with specified numerical values for the parameters.
186. **Structural Parameters:** Parameter which appear in the structural form equations..
187. **Structural form of a Simultaneous Equation System** : Representation of the system as a collection of autonomous behavioural, technological and accounting relationships.
188. **Smoothing**: Removing trends, cycles on seasonal variations from data.
189. **Technological relation:** A relationship usually cast aa an algebraic equation which describes a technological constraint.
190. **The t-value of a Coefficient**: Estimated coefficient divided by its estimated standard error. the statistic is used to test the hypothesis that the coefficient equals zero.
191. **The F-value of a Set of Coefficient**: The statistic used to test the hypothesis that all the coefficient in a set are simultaneously zero.
192. **Three Stage Least Squares(3SLS):** A procedure for joint estimation of all identified equations using the SUR procedure.
193. **Two Stage Least Squares (2SLS):** A procedure for the estimation of a single identified equation using two rounds of OLS.
194. **Unbiased Forecast**: Expected value of the forecast coincides with the expected values of the variable being forecast.
195. **Unconditional Forecast**: Forecasts in the presence of certain knowledge of values of explanatory variables.
196. **Union Of Two Events** : Given two events, their union is a third event which is said to have occurred when either or both of two events occur. union of more than two events is defined in the same manner.
197. **Unit non-response:** When a potential respondent does not provide data for any variables in a survey.
198. **Unbalanced panel:** A panel dataset that includes all respondents who report data for atleast one period (wave) of the panel. In contrast to a balanced panel which only includes those individuals with complete data for all periods.

199. **Variables:** Entities whose behaviour is being studied. generally they represent some measurable and observable economic construct.
200. **Variance Inflation Factors**: Diagonal element of covariance matrix of OLS estimators. they indicate the impact of multicollinearity on estimator variances.
201. **Variance of Random Variable** : A measure of dispersion around the mean in the values of a random variable. defined with respect to its density function.
202. **Von Neumann Ratio**: A statistic for testing for serial correlation in a series of random variables.
203. **Weibull model:** A parametric model for duration analysis.
204. **Weighted least squares:** Weights (wi) are attached to the values of the dependent variable (yi) and independent variables (xi) before using least squares regression. This method can be used to correct for heteroskedasticity.
205. **Zero Order Regression**: An estimation procedure for missing observations in which they are replaced by averages of the available observations.

****

(Question Pattern Based on MKU)
Sample Question Paper-1

TIME : Three hours MARKS: 75

PART-A

Answer the following questions (10X1=10)

1. **The term regression was coined by**
   A. Francis Galton B. Karlpearson
   C. Carl Friedrick Gauss.. D. William Sealy Goss
2. **Method of ordinary least square is attributed to**
   A. CarlFriedrick Gauss B.William Sealy Goss
   C.Durbin Watson D. Both b and c
3. **Locus of the conditional mean of the dependent variable for the fixed values of the explanatory variable**
   A. Indifference curve B. Population regression curve
   C. Production Possibility curve D. None of these
4. **In Yi= β1+β2X+ui, ui**
   A. Represent the missing values of Y
   B. Acts as proxy for all the omitted variables that may affect Y
   C. Acts as proxy for important variable that affect Y
   D. Represent measurement errors
5. **One of the assumption of CLRM is that the number of observations in the sample must be greater the number of**
   A. Regressor B. Regressands
   C. Dependent variable D. Dependent and independent variable
6. **The coefficient of determination shows,**
   A. Variation in the dependent variable Y is explained by the independent variable X
   B. Variation in the independent variable Y is explained by the dependent variable X.
   C. Both a and b are correct
   D. Both a and b are wrong
7. **What is the meaning of the term "heteroscedasticity"?**
   A. The variance of the errors is not constant
   B. The variance of the dependent variable is not constant
   C. The errors are not linearly independent of one another
   D. The errors have non-zero mean
8. **Near multicollinearity occurs when**
   A. Two or more explanatory variables are perfectly correlated with one another
   B. The explanatory variables are highly correlated with the error term
   C. The explanatory variables are highly correlated with the dependent variable
   D. Two or more explanatory variables are highly correlated with one another

9. If in our regression model, one of the explanatory variables included is the lagged value of the dependent variable, then the model is referred to as

A. Best fit model B. Dynamic model

C. Autoregressive model D. First-difference form

10. In binary logistic regression:

A. The dependent variable is continuous.

B. The dependent variable is divided into two equal subcategories.

C. The dependent variable consists of two categories.

D. There is no dependent variable.

PART-B

Answer the following questions either (a) Or (b) (5X7=35)

11. (a). Enumerate aim and objectives of Econometrics.

(Or)

(b).State significance of stochastic Error term

12. (a).Prove that $\hat{\alpha} = \overline{Y} - \overline{x}\frac{\Sigma x_i y_i}{\Sigma x_i^2}$ and $\hat{\beta} = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$ for a simple regression.

(Or)

(b). List out the properties of OLS estimators.

13. (a). Three related variables take following sets of values: Estimate a regression X_1 on X_2 and X_3

X_1: 123 4 5

X_2: 2 15 4 3

X_3: 3 14 5 2

(Or)

(b).Illustrate the application of Multiple Regression in day to day life.

14. (a) Write a note on Park test.

(Or)

(b) What are the consequences and remedies of Autocorrelation?

15. (a) What are the reasons for lags in econometrics?

(Or)

(b) Differentiates ANOVA and ANCOVA**.**

PART-C

Answer any one of the following questions (3X10=30)

16. Illustrate the methodology of Econometrics.

17. What are the assumptions of Classical Linear Regression Model?

18. Derive the formula for $\hat{\beta}$. of multiple regression in matrix form.

19. Enumerate the causes, consequences and remedies of Multicollinearity.

20. State the consequences of Model Specification Error.

****

Sample Question Paper-2

Time: 3 hrs **Max Marks: 75**

Answer all the Questions

PART-A **(10x1=10Marks)**

1. Econometric is an
   A) amalgamation of mathematics and statistics
   B) organization of Mathematics and Economics
   C) integration of Mathematics, Statistics and Economics
   D) integration of Economics and Statistics
2. One of the following assumption is not in OLS
   A) $E(u_i / x_i)= 0$ B) $Cov(u_iu_j/ x_ix_j) = 0$
   C) No Autocorrelation D) No Perfect Multicollinearity
3. In the simple linear regression model, the regression slope indicates
   (A) by how many percent *Y* increases, given a one percent increase in *X*.
   (B) the explanatory variable will give you the predicted *Y*.
   (C) by how many units *Y* increases, given a one unit increase in *X*.
   (D) Represents the elasticity of *Y* on *X*.
4. Find out which is not the property of a parameter?
   A) Linear in Parameter
   B) Parameters has the Minimum Variance
   C) Parameters are Unbiased
   D) Biased
5. The Co-efficient of Determination Measures
   A) The correlation between the X and Y
   B) Goodness of fit of the model
   C) Error
   D) TSS
6. The term Multiple regression Stands for
   A) Regressing more than one explanatory variables
   B) Regressing no variables
   C)Many regression
   D)Regressing one explanatory variable
7. Find out which is not the violation of assumption.
   A)Autocorrelation C) Multicollinearity
   B) Heteroscedasticity D) Dummy variable
8. Heteroscedasticity means that
   (A) homogeneity cannot be assumed automatically
   (B) the variance of the error term is not constant.

(C) the observed units have different preferences.
(D) agents are not all rational.

9. A binary variable is often called a
(A) dummy variable. (B) dependent variable.
(C) residual. (D) power of a test.

10. The difference between the Autoregressive and Distributed lag model is in its lag of
A) Dependent variable placing among explanatory variable
B) Dependent variable
C) Independent variable
D) Dummy variable.

PART-B (5x7=35)

11. a) Describe the objectives and features of Econometrics.

(OR)

b) Explain the Methodology of Econometrics.

12. a). Estimate the unknown parameter β_0 and β_1 in OLS estimation.

(OR)

b).Prove that the Property of Parameters are Unbiased.

13. a) What are assumptions of estimation of Multiple Regression?

(OR)

b) Derive the β value in terms of matrix as $(X'X)^{-1}X'Y$

14. a)Explain causes and consequence of Multicollinearity.

(OR)

b) Briefly explain reasons for Heteroscedasticity.

15. a).What are the reasons for the introduction of Lag in Regression?

(OR)

b). Find the difference between ANOVA and ANCOVA

PART-C (3x10=30)

Answer any Three questions:-

16. Illuminate the scope of Econometrics.
17. Enumerate the assumption of Classical Linear Regression model.
18. Three related variables take following sets of values: Estimate a regression X_1 on X_2 and X_3

X_1:	1	2	3	4	5
X_2:	2	1	5	4	3
X_3:	3	1	4	5	2

19. What is Autocorrelation? Explain the source, consequence and detection of Autocorrelation.
20. What are consequences of Model Specification Error?

****

Questions for Practice

1. Define Econometrics and State its objectives.
2. Enumerate the scope of Econometrics in daily routine life.
3. Explain the significance of U_i , in econometrics equations.
4. What are the differences between economic model and econometric model?
5. What are the features of Logit model?
6. What are the tests for detecting Heteroscedasticity?
7. What are the difference between analysis of variance and analysis of Co variance?
8. The estimated coefficient of different variables and its standard errors are given in the following table.. Find the't' values for each variable.

Variables	Co-efficient	Standard error
A_0	6.7213	0.7213
A_1	5.3121	0.3451
A_2	9.0932	0.7344
A_3	6.2436	0.5643
A_4	4.1256	0.3642

9. Illuminate the cautions in use of Dummy Variables?
10. Explain the methodology of Econometrics with help of Keynesian theory.
11. State the ten assumptions of Ordinary Least Square Method of estimation.
12. Analyse the causes and consequences of Multicollinearity and explain the remedial measures to eliminate the effects of Multicollinearity.
13. State and prove the Gauss Markov theorem.
14. Derive the $\beta_1$coefficient value for a multiple regression by using matrix method.
15. The following table includes the Price and Quantity demanded for Toffees:

Quantity (in numbers):	7	5	6	8	1	2
Price (Rs.):	2	4	3	1	6	5

    a) Estimate the demand function for Toffees $Y_i=\beta_0 +\beta_1 X_i + U_i$.
    b) Calculate the R^2
16. Prove that, in a given regression $Yi=\beta_0 +\beta_1 X_i + U_i$, the parameters β_0 and β_1 are linear and Unbiased.
17. Elucidate the Sources, Consequences and remedial measures to correct the problem of Autocorrelation and Heteroscedasticity.
18. Econometrics is a separate discipline. Why?

****$$$***

ALL THE BEST