M.COM SEMESTER - IV, (C. NO. M.COM-FE455)

Directorate of Distance & Online Education

UNIVERSITY OF JAMMU
JAMMU

SELF LEARNING MATERIAL

FOR

M.COM. FOURTH SEMESTER

FINANCIAL ECONOMETRICS

Session 2024-2025

COURSE NO: M.COM-FE455	**FINANCE & ACCOUNTING GROUP**	UNIT : I-IV

Course Coordinator :

Prof. Gurjeet Kour

Head, Department of Commerce

University of Jammu

http:/www.distanceeducation.in

Printed and published on behalf of the Directorate of Distance & Online Education, University of Jammu, Jammu by the Director, DD&OE, University of Jammu, Jammu

COURSE NO: M.COM-FE455

Written by :
Dr. Pinkey
Assistant Professor, Commerce
Govt. Degree College, Boys, Kathua
Unit I & II

Dr. Kanchan Kumari
Lecturer, Commerce
Govt. Degree College, Akhnoor
Unit III & IV

Reviewed & Edited by :
Dr. Avantika Bakshi
Teacher Incharge, Commerce
DD&OE, University of Jammu.

Printed at :- **SNEH PRINTERS** /2024/700 Qty.

SYLLABUS

M.COM. FOURTH SEMESTER (NON CBCS)

FINANCIAL ECONOMETRICS

FINANCE AND ACCOUNTING GROUP

Course: M.COM FE455 **Max Marks: 100 Marks**

Credit: 4 **External: 80 Marks**

Time: 3.00 Hrs **Internal: 20 Marks**

(Syllabus for the examination to be held in May2024, 2025, 2026)

COURSE OBJECTIVES

1. To make the students familiar with the basic concepts of econometrics.
2. To sensitize the students about the postulation and testing of hypotheses pertaining toeconomic issues or problems.
3. To provide the students a thorough grounding of the basic econometric techniques andtheir applications and equip them with a broad knowledge of regression analysis relevant foranalysing economic data.
4. To build confidence among students to interpret and critically evaluate the outcomes of empirical analysis by using standard econometrics techniques.

COURSE OUTCOMES

After the completion of this course, the students will be able to:

1. have a deeper understanding of basic concepts of econometrics and have greater confidence in its application;
2. interpret and critically evaluate the outcomes of empirical analysis by using standard econometrics techniques;
3. learn appropriate alternatives to ordinary least squares, when assumptions underlying the classical linear regression model are violated.

4. use a statistical/econometric computer package to estimate an econometric model in computer software like EVIEWS, STATA and Gretel; and
5. act as a professional economist for the government, nongovernment and the corporate sectors.

UNIT - I INTRODUCTION

Basics of econometrics; Scope of econometrics; Methodology of econometrics; Types ofeconometrics; Difference between econometrics and statistics; Nature and source of data used for economic analysis, the accuracy of the data; Basics of regression; Two variable regressionmodel-assumptions, estimation through OLS; Properties of estimates-Gauss MarkovTheorem;Concept and derivation of R2 and adjusted R2; Deviation from classical linear regressionassumptions and GLS.

UNIT II PROBLEM WITH REGRESSION ANALYSIS

Problem with regression analysis- problem of heteroskedasticity-nature, test, consequencesand remedial measures; Problem of autocorrelation-nature, test, consequences and remedialmeasures; Problem of Multicollinearity- nature, test, consequences and remedial measures;Model mis-specification verses pure autocorrelation, OLS verses FGLS and HAC; Co-existence of autocorrelation and heteroskedasticity.

UNIT - III REGRESSION WITH QUALITATIVE VARIABLE

Dummy variables- Basics, testing structural stability of regression models; Dummy variable trap, basics of trap, regression with dummy dependent variables; LMP Model-logit, grouped logit, probit and tobit model- their applications; Modelling count data, poission model.

UNIT IV TIME SERIES ECONOMETRICS

Time series analysis-Basics of time series; Utility of time series; Components of time seriessecular trend, seasonal variations, cyclical variations, irregular variations, preliminaryadjustments before analysing time series; Time series

econometrics-Stochastic processes; Stationary stochastic processes; Non stationary stochastic processes; Random walk models; Cointegration; Deterministic and stochastic trends; Unit root tests; Approaches to economic forecasting, AR, MA and ARIMA modelling of time series data.

BOOKS RECOMMENDED:

1. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
2. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
3. Baltagi, B. Basic Econometrics. Springer, New Delhi.
4. Ramu, R. Introductory Econometrics with Applications. South-Western College PublishingCompany, USA.

Note: Latest edition of the books may be preferred.

NOTE FOR PAPER SETTING

The paper consists of two sections. Each section will cover the whole of the syllabus without repeating the question in the entire paper.

Section A: It will consist of eight short answer questions, selecting two from each unit. A candidate must attempt any six and answer to each question shall be within 200 words. Each question carries four marks and total weightage to this section shall be 24 marks.

Section B: It will consist of six essay type questions with answer to each question within 800 words. One question will be set atleast from each unit and the candidate must attempt four. Each question will carry 14 marks and total weightage shall be 56 marks.

MODEL QUESTION PAPER

SECTION-A

Time : 3 hrs **Max Marks : 80**

Note:- Attempt any six questions. Each question carries 4 marks. Answer to each question should be within 200 words.

1. Explain stages of development of econometrics?
2. Explain the scope and types of econometrics?
3. Explain time series and Cross-section data?
4. What is pure Autocorrelation?
5. Explain the meaning of derivation of R2?
6. Explain the nature and testing of heteroskedasticity?
7. Explain Durbin Watson Test.
8. Explain the method of Generalized Least Squares.

SECTION-B

Note:- Attempt any four questions. Each question carries 14 marks. Answer to each question should be within 800 words.

1. Explain Econometric Models?
2. Explain theoretical and applied econometrics?
3. Explain Gauss-Markov theorem.
4. Explain the assumptions of classical linear regression models?
5. Explain the consequences and various remedial measures for the problem of heteroskedasticity.
6. Explain the consequences and various remedial measures for the problem of autocorrelation?
7. Explain the consequences and various remedial measures for the problem of multicollinearity?
8. Explain the concept of coexistence of autocorrelation and heteroscedasticity ?

CONTENTS

UNIT-I **LESSON NO. 1**

INTRODUCTION OF FINANCIAL ECONOMETRICS

BASICS OF ECONOMETRICS

STRUCTURE

1.1 INTRODUCTION

Econometrics deals with the measurement of economic relationships. It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are usually expressed in mathematical forms and combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships. The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods. The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations. It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

It may be pointed out that the econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc. In simple words, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help. The econometric tools are helpful in explaining the relationships among variables.

Econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet the subject deserves to be studied for the following reasons. Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity. But the

theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell by how much the quantity will go up or down because of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Stated differently, econometrics gives empirical content to most economic theory. The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. Econometrics, as noted previously, is mainly interested in the empirical verification of economic theory. As we shall see, the econometrician often uses the mathematical equations proposed by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. And this conversion of mathematical into econometric equations requires a great deal of ingenuity and practical skill. Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. These are the jobs of the economic statistician. It is he or she who is primarily responsible for collecting data on gross national product (GNP), employment, unemployment, prices, and so on. The data thus collected constitute the raw data for econometric work. But the economic statistician does not go any further, not being concerned with using the collected data to test economic theories. Of course, one who does that becomes an econometrician. Although mathematical statistics provides many tools used in the trade, the econometrician often needs special methods in view of the unique nature of most economic data, namely, that the data are not generated as the result of a controlled experiment. The econometrician, like the meteorologist, generally depends on data that cannot be controlled directly.

The term "econometrics" is believed to have been crafted by **Ragnar Frisch (1895-1973)** of Norway, one of the three principal founders of the Econometric Society, first editor of the journal *Econometrica,* and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch's own words in the introduction to the first issue of *Econometrica* to describe the discipline.

1.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- the concept of econometrics,
- stages of development,
- testing the hypothesis,
- econometric models,
- aims of econometrics,
- relation in econometrics and statistics,
- types of data,
- aggregation problem,
- econometrics and regression analysis and
- linear regression model.

1.3 BASICS OF ECONOMETRICS

1.3.1 Concept of Econometrics

Literally interpreted, econometrics means "economic measurement." Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations:

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

Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.

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.

Econometrics is concerned with the empirical determination of economic laws.

The art of the econometrician consists in finding the set of assumptions that are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him.

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

Today, we would say that econometrics is the unified study of economic models, mathematical statistics, and economic data. Within the field of econometrics there are sub-divisions and specializations. **Econometric theory** concerns the development of tools and methods, and the study of the properties of econometric methods. **Applied econometrics** is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

Econometrics is based upon the development of statistical methods for estimating economic relationships, testing economic theories, and evaluating and implementing government and business policy. The most common application of econometrics is the forecasting of such important macroeconomic variables as interest rates, inflation rates, and gross domestic product. Whereas forecasts of economic indicators are highly visible and often widely published, econometric methods can be used in economic areas that have nothing to do with macroeconomic forecasting.

Economists develop economic models to explain consistently recurring relationships. Their models link one or more economic variables to other economic variables. For example, economists connect the amount individuals spend on consumer goods to disposable income and wealth and expect consumption to increase as disposable income and wealth increase (that is, the relationship is positive).

There are often competing models capable of explaining the same recurring relationship, called an empirical regularity, but few models provide

useful clues to the magnitude of the association. Yet this is what matters most to policymakers. When setting monetary policy, for example, central bankers need to know the likely impact of changes in official interest rates on inflation and the growth rate of the economy. It is in cases like this that economists turn to econometrics.

Econometrics uses economic theory, mathematics, and statistical inference to quantify economic phenomena. In other words, it turns theoretical economic models into useful tools for economic policymaking. The objective of econometrics is to convert qualitative statements (such as "the relationship between two or more variables is positive") into quantitative statements (such as "consumption expenditure increases by 95 cents for every one dollar increase in disposable income").

Econometricians, practitioners of econometrics, transform models developed by economic theorists into versions that can be estimated. As Stock and Watson (2007) put it, "econometric methods are used in many branches of economics, including finance, labor economics, macroeconomics, microeconomics, and economic policy." Economic policy decisions are rarely made without econometric analysis to assess their impact.

The main tool of econometrics is the linear multiple regression model, which provides a formal approach to estimating how a change in one economic variable, the explanatory variable, affects the variable being explained, the dependent variable- taking into account the impact of all the other determinants of the dependent variable. This qualification is important because a regression seeks to estimate the marginal impact of a particular explanatory variable after considering the impact of the other explanatory variables in the model. For example, the model may try to isolate the effect of a one percentage point increase in taxes on average household consumption expenditure, holding constant other determinants of consumption, such as pretax income, wealth, and interest rates.

1.3.2 Stages of development

The methodology of econometrics is straight forward.The first step is

to suggest a theory or hypothesis to explain the data being examined. The explanatory variables in the model are specified, and the sign and/or magnitude of the relationship between each explanatory variable and the dependent variable are clearly stated. At this stage of the analysis, applied econometricians rely heavily on economic theory to formulate the hypothesis. For example, a tenet of international economics is that prices across open borders move together after allowing for nominal exchange rate movements (purchasing power parity). The empirical relationship between domestic prices and foreign prices (adjusted for nominal exchange rate movements) should be positive, and they should move together approximately one for one.

The second step is the specification of a statistical model that captures the essence of the theory the economist is testing. The model proposes a specific mathematical relationship between the dependent variable and the explanatory variables - on which, unfortunately, economic theory is usually silent. By far the most common approach is to assume linearity -meaning that any change in an explanatory variable will always produce the same change in the dependent variable (that is, a straight-line relationship).

Because it is impossible to account for every influence on the dependent variable, a catchall variable is added to the statistical model to complete its specification. The role of the catchall is to represent all the determinants of the dependent variable that cannot be accounted for - because of either the complexity of the data or its absence. Economists usually assume that this "error" term averages to zero and is unpredictable, simply to be consistent with the premise that the statistical model accounts for all the important explanatory variables.

The third step involves using an appropriate statistical procedure and an econometric software package to estimate the unknown parameters (coefficients) of the model using economic data. This is often the easiest part of the analysis thanks to readily available economic data and excellent econometric software. Still, the famous GIGO (garbage in, garbage out) principle of computing also applies to econometrics. Just because something can be computed doesn't mean it makes economic sense to do so.

The fourth step is by far the most important: administering the smell test. Does the estimated model make economic sense -that is, yield meaningful economic predictions? For example, are the signs of the estimated parameters that connect the dependent variable to the explanatory variables consistent with the predictions of the underlying economic theory? (In the household consumption example, for instance, the validity of the statistical model would be in question if it predicted a decline in consumer spending when income increased). If the estimated parameters do not make sense, how should the econometrician change the statistical model to yield sensible estimates? And does a more sensible estimate imply an economically significant effect? This step calls on and tests the applied econometrician's skill and experience.

1.3.3 Testing the hypothesis

The main tool of the fourth stage is hypothesis testing, a formal statistical procedure during which the researcher makes a specific statement about the true value of an economic parameter, and a statistical test determines whether the estimated parameter is consistent with that hypothesis. If it is not, the researcher must either reject the hypothesis or make new specifications in the statistical model and start over.

If all four stages proceed well, the result is a tool that can be used to assess the empirical validity of an abstract economic model. The empirical model may also be used to construct a way to forecast the dependent variable, potentially helping policymakers make decisions about changes in monetary and/or fiscal policy to keep the economy on an even keel.

Students of econometrics are often fascinated by the ability of linear multiple regression to estimate economic relationships. Three fundamentals of econometrics are worth remembering.

- First, the quality of the parameter estimates depends on the validity of the underlying economic model.
- Second, if a relevant explanatory variable is excluded, the most likely outcome is poor parameter estimates.

- Third, even if the econometrician identifies the process that actually generated the data, the parameter estimates have only a slim chance of being equal to the actual parameter values that generated the data. Nevertheless, the estimates will be used because, statistically speaking, they will become precise as more data become available.

Econometrics, by design, can yield correct predictions on average, but only with the help of sound economics to guide the specification of the empirical model. Even though it is a science, with well-established rules and procedures for fitting models to economic data, in practice econometrics is an art that requires considerable judgment to obtain estimates useful for policymaking.

1.3.4 Econometric Models

A model is a simplified representation of a real-world process. It should be representative in the sense that it should contain the salient features of the phenomena under study. In general, one of the objectives in modeling is to have a simple model to explain a complex phenomenon. Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic. In practice, generally, all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model. Rest of the variables are dumped in a basket called "disturbances" where the disturbances are random variables. This is the main difference between economic modeling and econometric modeling. This is also the main difference between mathematical modeling and statistical modeling. The mathematical modeling is exact in nature, whereas the statistical modeling contains a stochastic term also. An economic model is a set of assumptions that describes the behaviour of an economy, or more generally, a phenomenon.

An econometric model consists of:

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

1.3.5 Aims of Econometrics

The three main aims econometrics are as follows:

1. **Formulation and specification of econometric models:** The economic models are formulated in an empirically testable form. Several econometric models can be derived from an economic model. Such models differ due to different choice of functional form, specification of the stochastic structure of the variables etc.

2. **Estimation and testing of models:** The models are estimated based on the observed set of data and are tested for their suitability. This is the part of the statistical inference of the modelling. Various estimation procedures are used to know the numerical values of the unknown parameters of the model. Based on various formulations of statistical models, a suitable and appropriate model is selected.

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

1.3.6 Econometrics and statistics

Econometrics differs both from mathematical statistics and economic statistics. In economic statistics, the empirical data is collected recorded, tabulated, and used in describing the pattern in their development over time. The economic statistics is a descriptive aspect of economics. It does not provide either the explanations of the development of various variables or measurement of the parameters of the relationships. Statistical methods describe the methods of measurement which are developed on the basis of controlled experiments. Such methods may not be suitable for the economic phenomenon as they don't fit in the framework of controlled experiments. For example, in real-world experiments, the variables usually change continuously and simultaneously, and so the set up of controlled experiments are not suitable.

Econometrics uses statistical methods after adapting them to the

problems of economic life. These adopted statistical methods are usually termed as econometric methods. Such methods are adjusted so that they become appropriate for the measurement of stochastic relationships. These adjustments basically attempt to specify attempts to the stochastic element which operate in real-world data and enters the determination of observed data. This enables the data to be called a random sample which is needed for the application of statistical tools.

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

The **applied econometrics** includes the application of econometric methods to specific branches of econometric theory and problems like demand, supply, production, investment, consumption etc. The applied econometrics involves the application of the tools of econometric theory for the analysis of the economic phenomenon and forecasting economic behaviour.

1.3.7 Types of data

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

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

2. **Cross-section data**: The cross-section data give information on the variables concerning individual agents (e.g., consumers or producers) at a given point of time. For example, a cross-section of a sample of consumers is a sample of family budgets showing expenditures on various commodities by each family, as well as information on family income, family composition and other demographic, social or financial characteristics.

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

4. **Dummy variable data:** When the variables are qualitative in nature, then the data is recorded in the form of the indicator function. The values of the variables do not reflect the magnitude of the data. They reflect only the presence/absence of a characteristic. For example, variables like religion, gender, taste, etc. are qualitative variables. The variable 'gender' takes two values - male or female, the variable 'taste' takes values - like or dislike etc. Such values are denoted by the dummy variable. For example, these values can be represented as '1' represents male and '0' represents female. Similarly, '1' represents the liking of taste, and '0' represents the disliking of taste.

1.3.8 Aggregation problem

The aggregation problems arise when aggregative variables are used in functions. Such aggregative variables may involve.

1. **Aggregation over individuals:** For example, the total income may comprise the sum of individual incomes.
2. **Aggregation over commodities:** The quantity of various commodities may be aggregated over, e.g., price or group of commodities. This is done by using suitable index.
3. **Aggregation over time periods**: Sometimes the data is available for shorter or longer time periods than required to be used in the functional form of the economic relationship. In such cases, the data needs to be aggregated over the time. For example, the production of most of the manufacturing commodities is completed in a period shorter than a year. If annual figures are to be used in the model, then there may be some error in the production function.
4. **Spatial aggregation:** Sometimes the aggregation is related to spatial issues. For example, the population of towns, countries, or the production in a city or region etc. Such sources of aggregation introduce "aggregation bias" in the estimates of the coefficients. It is important to examine the possibility of such errors before estimating the model.

1.3.9 Econometrics and Regression Analysis:

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

1.3.10 Linear regression model:

Suppose the outcome of any process is denoted by a random variable y, called as dependent (or study) variable, depends on k independent (or explanatory) variables denoted by $X_1 X_{2,\ldots\ldots} X_k$. Suppose the behaviour of y can be explained by a relationship given by

$$y = \int (X_1 X_3 \ldots . X_3 \ldots . X_{10} \beta_1 \beta_2, \ldots . , \beta_k) + \varepsilon$$

where $\int$ is some well-defined function and $\beta_1 \beta_2, \ldots . , \beta_k$ are the parameters which characterize the role and contribution of X_1 X_2,….. X_k, respectively. The term reflects the stochastic nature of the relationship between y and X_1 X_2,….. X_k and indicates that such a relationship is not exact in nature. When $\varepsilon = 0$, then the relationship is called the mathematical model otherwise the statistical model. The term "model" is broadly used to represent any phenomenon in a mathematical framework.

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

For example: $Y = \beta_{1,} X_1^2 + \beta_2 \varepsilon$

1.4 SUMMARY

In last, we can say that the subject of econometrics deals with the economic measurement. And further, it is defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena. It is also concerned with the empirical determination of economic law.

1.5 GLOSSARY

- **Parameter:** Something that decides or limits the way in which something can be done
- **Linear model:** Linear regression is a statistical method used to create a linear model. The model describes the relationship between a dependent variable y (also called the response) as a function of one or more independent variables Xi (called the predictors).
- **Theoretical econometrics:** Theoretical econometricians investigate the properties of existing statistical tests and procedures for estimating unknowns in the model. They also seek to develop new statistical procedures that are valid (or robust) despite the peculiarities of economic datasuch as their tendency to change simultaneously.
- **Applied econometrics:** Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analysing economic history, and forecasting.
- **Time series data:** Time series data is a collection of observations obtained through repeated measurements over time.
- **Cross-section data:** Cross-sectional data, or a cross section of a study population, in statistics and econometrics, is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time. The analysis might also have no regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among selected subjects.

- **Panel data:** Panel data, sometimes referred to as longitudinal data, is data that contains observations about different cross sections across time. Examples of groups that may make up panel data series include countries, firms, individuals, or demographic groups.
- **Dummy variable data:** In statistics and econometrics, particularly in regression analysis, a dummy variable is one that takes only the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome.
- **Catchall variable:** Catchall variable is a term or category which includes many different things.
- **Linear regression:** Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).

1.6 SELF-ASSESSMENT QUESTIONS

Q1. What are the fundamentals of econometrics?

_____

_____

_____

Q2. Define theoretical econometrics?

_____

_____

_____

Q3. Define applied econometrics?

_____

_____

_____

Q4. Briefly explain time series data?

_____

_____

_____

Q5. Explain cross section data?

_____

_____

_____

Q6. What is Panel data?

_____

_____

_____

Q7. Explain dummy variable data?

_____

_____

_____

1.7 LESSON END EXERCISE

Q1. Explain the concept of econometrics?

_____

_____

_____

Q2. Explain stages of development of econometrics?

_____

_____

_____

Q3. Explain testing of hypothesis in econometrics?

_____

_____

_____

Q4. Explain Econometric Models?

_____

_____

_____

Q5. Explain aims of econometrics?

_____

_____

_____

Q6. Explain the relation between econometrics and statistics?

_____

_____

_____

Q7. Explain various types of data in econometrics?

_____

_____

_____

Q8. Explain aggregation problem in econometrics?

_____

_____

_____

Q9. Explain Linear regression model in econometrics?

_____

_____

_____

1.8 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.
10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.
11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-I **LESSON NO. 2**

INTRODUCTION OF FINANCIAL ECONOMETRICS

SCOPE, METHODOLOGY, TYPES OF ECONOMETRICS

STRUCTURE

2.1 INTRODUCTION

Econometrics refers to the application of economic theory and statistical techniques for the purpose of testing hypothesis and estimating and forecasting economic phenomenon. Literally interpreted, econometrics means "economic measurement." Although measurement is animportant part of econometrics,

the scope of econometrics is much broader, as can be seen from the following quotations: Econometrics, the result of a certain outlook on the role ofeconomics, consists of the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results. econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference. Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis economic phenomena.Econometrics is concerned with the empirical determination of economic laws.

2.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- the scope of econometrics,
- methodology of econometrics,
- types of econometrics and
- difference between econometrics and statistics

2.3 SCOPE OF ECONOMETRICS

Econometrics is the application of statistical methods and mathematics to economic data. It is a branch of economics that focuses on giving experimental content for finding out economic relations. It also aims at computing relationships between economic variables through statistical techniques. There are many courses available in the field of Econometrics. Some of them are Graduate Diploma in Econometrics, B.A in Economics with Econometrics, M.Sc in Econometrics, M.A in Econometrics, etc.

Quantitative economics is a highly specialized field of study which is taught at the post-graduation level. The courses related to this field are popular in India, but it is done by some of the best brains of the society. The study of quantitative economics became more important because of the utilization of

economics as a subject that can analytically approach the problems and provide efficient solutions. The subject is also known as econometrics as it deals with the study of complex mathematical and statistical models which help in detailed study of concepts of economics.

The growth and development of industries is always dependent on a few factors such as resource utilization, maximization of revenue and similar factors. The subject quantitative economics provide economic models required for the analysis of such factors. The demand for experts in this field are huge across all the sectors some of them being advisory bodies, multinational corporations, manufacturing units, business conglomerates and so on.

The global competition has led to unending race where every enterprise wants to become the market leader. This has led to the enhanced quality and services provided by organizations. The role of economic models in such a scenario becomes even more important. Econometrics is a highly important subject for research purpose because it always leads to an efficient solution of an economic problem.

Several esteemed institutes of the country offer course on this subject. In a developing economy like India, the role of econometrists becomes obligatory. The experts in this field are the ones who bag the best jobs in the market. Not only in India but the employment scope for the quantitative economics experts is huge abroad also.

2.4 METHODOLOGY OF ECONOMETRICS

It means how does the econometrician go ahead in analysing an economic theory. What is needed is a methodology, i.e. a step-by-step procedure. This is like other social sciences. A theory should have a prediction. In statistics and econometrics, we also speak of **hypothesis**. One example is **the marginal propensity to consume** (MPC) proposed by Keynes. Other examples could be that lower taxes would increase growth, or maybe that it would increase economic inequality, and that introducing a common currency has a positive effect on trade. Below mentioned points briefly describes the methodology of econometrics:

(i) Specification of the Mathematical Model: This is where the algebra enters. We need to use mathematical skills to produce an equation. Assume a theory predicting that more schooling increases the wage. In economic terms, we say that the return to schooling is positive. The equation is:

$$Y = \beta_1 + \beta_2 X,$$

where Y is the variable for wage and β_1 is a constant and β_2 is the coefficient of schooling, and X is a measurement of schooling, i.e. the number of years in school. We also call β_1 intercept and β_2 a slope coefficient.

Normally, we would expect both β_1 and β_2 to be positive.

(ii) Specification of the Econometric Model: Here, we assume that the mathematical model is correct, but we need to account for the fact that it may not be so. We add an error term, u to the equation above. It is also called a random (stochastic) variable. It represents other non-quantifiable or unknown factors that affect Y. It also represents mismeasurements that may have entered the data. The econometric equation is:

$$Y = \beta_1 + \beta_2 X + u.$$

The error term is assumed to follow some sort of statistical distribution. This will be important later.

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

(iv) Estimation of the model: Here, we quantify β_1 and β_2, i.e. we obtain numerical estimates. This is done by statistical technique called **regression analysis.**

(v) Hypothesis Testing: Now we go back to the part where we had economic theory. The prediction was that schooling is good for the wage. Now we check does the econometric model support this hypothesis. What we do here is called **statistical inference (hypothesis testing)**. Technically speaking,

the β_2 coefficient should be greater than 0 if schooling is positively effect on wages (accept the hypothesis).

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

(vii) Use for Policy Recommendation: Lastly, if the theory seems to make sense and the econometric model was not refuted based on the hypothesis test, we can go on to use the theory for policy recommendation. If your theory was good, then maybe you will earn the Nobel Prize of Economics.

2.5 TYPES OF ECONOMETRICS

There are two branches of econometrics: theoretical econometrics and applied econometrics.

(i) Theoretical Econometrics: It is the study of the properties of existing statistical models and procedures for finding out the unknown values in the model. In this we seek to develop new statistical procedures that are valid despite the nature of economic data to change itself simultaneously.

Theoretical econometrics relies heavily on the likes of mathematics, theoretical statistics, and numerical quantities to prove that the new procedures have the ability to draw correct inferences.

The theoretical econometrics focuses on issues such as the general linear model, simultaneous equations models, distributed lags and ancillary related topics. Most of these problems were encountered while working on empirical research.

(ii) Applied Econometrics: It is the special use of econometric techniques to convert qualitative economic statements into quantitative ones, unlike the theoretical approach. Because applied econometricians acquire a closer experience with the data, they often face problems regarding data

attributes that point to errors with existing set of estimation techniques and alert their theoretical econometricians about the anomalies.

The applied econometrics deals with topics of production of goods and their productivity, demand for labour, arbitrage pricing theory, demand for housing related issues.

For example, the econometrician might discover that the variance of the data (how much individual values in a series differ from the overall average) is always rotating and is never fixed over time.

2.6 DIFFERENCE BETWEEN ECONOMETRICS AND STATISTICS

Econometrics and statistics have common overlapping areas that some people may find confusing. While both fields deal with statistics and the relationship between data, they are different. Before learning how these two differ from each other, it is crucial to understand what they are.

Econometrics: The term econometrics is a combination of two words, "econ" and "metrics". "Econ" refers to economics, social science that studies the production, distribution, and consumption of goods and services. "Metrics" means a system or standard of measurement. Econometrics is a field within economics that involves the quantification of economic data.

Econometrics uses statistical and mathematical models to analyze economic theories. This process has a crucial application within economics. Similarly, through econometrics, analysts can test and develop economic theories. They can also use the information in predictive modeling. For example, analysts can create time series models using the application of econometrics.

Econometrics includes three primary areas. These include mathematics, statistics, and economic theory. However, econometrics is not the same as mathematical science, economic statistics, or general economic theory. Instead, it combines all of these to help analysts understand the quantitative relations in modern economic life.

Statistics: Statistics is a much broader concept compared to econometrics. It is the branch of applied mathematics that involves collecting, reviewing, analysing, and inferring conclusions from quantitative data. The application of statistics is prevalent in almost every field, particularly scientific. This field generally focuses on two areas, uncertainty, and variation.

The primary objective of applying statistics is to draw a conclusion about a large number of events based on observable characteristics of small samples. There are two significant areas within statistics, known as descriptive and inferential statistics. Descriptive statistics involves describing the properties and sample and population data. Inferential statistics, on the other hand, deals with testing theories and reaching conclusions.

There are several tools within statistics that statisticians use. These may include variance, skewness, kurtosis, analysis of variance, null hypothesis testing, etc. Some of these tools may also have application in econometrics, such as regression analysis. Apart from other fields, economists also use statistics to collect, review and analyse data. Based on this information, they can draw conclusions, which is also a part of statistics.

Difference between Statistics and Econometrics:

The difference between statistics and econometrics comes from their fundamental areas of study. Statistics primarily relates to applied mathematics. Econometrics, on the other hand, is a part of economics. On top of that, statistics covers a significantly large area of study. While econometrics also includes statistics, it is not as broad.

Econometrics depends on statistics and statistical models to work. However, it doesn't only include these. It also consists of mathematics and economic theory, both of which are a fundamental part of it. The statistics used in econometrics only involves a particular area of the field. On top of that, econometrics includes other areas, such as causal inference and time series. These areas, while included in statistics, are not as prominent in the field.

2.7 SUMMARY

Econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity. But the theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell by how much the quantity will go up or down as a result of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Stated differently, econometrics gives empirical content to most economic theory. The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory. Econometrics is mainly interested in the empirical verification of economic theory. Econometricianoften uses the mathematical equations proposed by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. And this conversion of mathematical into econometric equations requires a great deal of ingenuity and practical skill.

2.8 GLOSSARY

- **Quantitative economics:** Quantitative Economics is the study of how we use our resources for the production, distribution, and consumption of goods and services. Economists study problems such as inflation and unemployment.
- **Marginal propensity to consume (MPC):** The marginal propensity to consume is equal to AC / AY, where AC is the change in consumption, and AY is the change in income. If consumption increases by 80 cents for each additional dollar of income, then MPC is equal to 0.8 / 1 = 0.8.

- **Econometric Model:** Econometric models are constructed from economic data with the aid of the techniques of statistical inference. These models are usually based on economic theories that assume optimizing behavior on the part of economic agents.
- **Random (stochastic) variable:** A random variable (stochastic variable) is a type of variable in statistics whose possible values depend on the outcomes of a certain random phenomenon. Since a random variable can take on different values, it is commonly labeled with a letter (e.g., variable "X").
- **Regression analysis:** Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.
- **Statistical inference:** Statistical inference is the process of drawing conclusions about populations or scientific truths from data. There are many modes of performing inference including statistical modeling, data-oriented strategies and explicit use of designs and randomization in analyses.
- **Hypothesis testing:** Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution.
- **Theoretical econometricians:** Theoretical econometricians investigate the properties of existing statistical tests and procedures for estimating unknowns in the model. They also seek to develop new statistical procedures that are valid (or robust) despite the peculiarities of economic data such as their tendency to change simultaneously.
- **Applied econometrics:** Applied econometrics uses theoretical econometrics and real-world data for assessing economic theories, developing econometric models, analysing economic history, and forecasting.

- **Statistics:** Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.

2.9 SELF-ASSESSMENT QUESTIONS

Q1. What is quantitative economics?

_____

_____

_____

Q2. Define regression analysis?

_____

_____

_____

Q3. What is marginal propensity to consume?

_____

_____

_____

Q4. Briefly explain random (stochastic) variable?

_____

_____

_____

Q5. Explain theoretical econometrics?

_____

_____

_____

Q6. What is applied econometrics?

_____

_____

_____

2.10 LESSON END EXERCISE

Q1. Explain the scope of econometrics?

_____

_____

_____

Q2. Explain types of econometrics?

_____

_____

_____

Q3. Explain methodologyofeconometrics?

_____

_____

_____

Q4. Differentiate between econometrics and statistics?

_____

_____

_____

2.11 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.

4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.

5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.

7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.

8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.

9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.

10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

******

UNIT-I **LESSON NO. 3**

INTRODUCTION OF FINANCIAL ECONOMETRICS

NATURE AND SOURCE OF DATA USED FOR ECONOMIC ANALYSIS, ACCURACY OF DATA

STRUCTURE

3.1 INTRODUCTION

Economic data are data describing an actual economy, past or present. These are typically found in time-series form, that is, covering more than one time period (say the monthly unemployment rate for the last five years) or in cross-sectional data in one time period (say for consumption and income levels for sample households). Data may also be collected from surveys of for example

individuals and firms or aggregated to sectors and industries of a single economy or for the international economy. A collection of such data in table form comprises a data set.

Methodological economic and statistical elements of the subject include measurement, collection, analysis, and publication of data. 'Economic statistics'may also refer to a subtopic of official statistics produced by official organizations (e.g. statistical institutes, intergovernmental organizations such as United Nations, European Union or OECD, central banks, ministries, etc.).

Good economic data are a precondition to effective macroeconomic management. With the complexity of modern economies and the lags inherent in macroeconomic policy instruments, a country must have the capacity to promptly identify any adverse trends in its economy and to apply the appropriate corrective measure. This cannot be done without economic data that is complete, accurate and timely.

Increasingly, the availability of good economic data is coming to be seen by international markets as an indicator of a country that is a promising destination for foreign investment. International investors are aware that good economic data is necessary for a country to effectively manage its affairs and, other things being equal, will tend to avoid countries that do not publish such data.

The public availability of reliable and up-to-date economic data also reassures international investors by allowing them to monitor economic developments and to manage their investment risk. The severity of the Mexican and Asian financial crises was made worse by the realization by investors that the authorities had hidden a deteriorating economic situation by slow and incomplete reporting of critical economic data. Being unsure of exactly how bad the economic situation was, they tried to withdraw their assets quickly and in the process caused further damage to the economies in question. It was the realization that data issues lay behind much of the damage done by these international financial crises that led to the creation of international data quality standards, such as the International Monetary Fund (IMF) General Data Dissemination System (GDDS).

Inside a country, the public availability of good quality economic data allows firms and individuals to make their business decisions with confidence that they understand the overall macroeconomic environment. As with international investors, local business people are less likely to overreact to a piece of bad news if they understand the economic context.

Tax data can be a source of economic data. In the United States, the IRS provides tax statistics, but the data are limited by statutory limitations and confidentiality concerns.

3.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- the nature of data used in economic analysis,
- sources of data used for economic analysis and
- the accuracy of data.

3.3 NATURE OF DATA USED IN ECONOMIC ANALYSIS

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

1. **Time series data:** Time series data give information about the numerical values of variables from period to period and are collected over time. For example, the data during the years 1990-2010 for monthly income constitutes a time series of data. A time series is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals, such as daily (e.g., stock prices, weather reports), weekly (e.g., money supply figures), monthly (e.g., the unemployment rate, the Consumer Price Index [CPI]), quarterly (e.g., GDP), annually (e.g., government budgets), quinquennially, that is, every 5 years (e.g., the census of manufactures), or decennially, that is, every 10 years (e.g., the census of population). Sometime data are available both quarterly as well as annually, as in the case of the data on GDP and consumer expenditure. With the advent of high-speed computers, data can now be collected over an extremely short interval of time, such as

the data on stock prices, which can be obtained literally continuously (the so-called real-time quote).

Although time series data are used heavily in econometric studies, they present special problems for econometricians. As we will show in chapters on time series econometrics later, most empirical work based on time series data assumes that the underlying time series is stationary. Although it is too early to introduce the precise technical meaning of stationarity at this juncture, loosely speaking, a time series is stationary if its mean and variance do not vary systematically over time.

2. **Cross-section data:** Cross-section data are data on one or more variables collected at the same point in time, such as the census of population conducted by the Census Bureau every 10 years (the latest being in year 2000), the surveys of consumer expenditures conducted by the University of Michigan, and, of course, the opinion polls by Gallup and umpteen other organizations.

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

3. **Panel, Longitudinal, or Micropanel data:** The panel data are the data from a repeated survey of a single (cross-section) sample in different periods of time. This is a special type of pooled data in which the same cross-sectional unit (say, a family or a firm) is surveyed over time. For example, the U.S. Department of Commerce carries out a census of housing at periodic intervals. At each periodic survey the same household (or the people living at the same address) is interviewed to find out if there has been any change in the housing and financial conditions of that household since the last survey. By interviewing the same household periodically, the panel data provide very useful information on the dynamics of household behaviour.

4. **Dummy variable data:** When the variables are qualitative in nature, then the data is recorded in the form of the indicator function. The values of the variables do not reflect the magnitude of the data. They reflect only the presence/absence of a characteristic. For example, variables like religion, sex, taste, etc. are qualitative variables. The variable 'sex' takes two values – male or female, the variable 'taste' takes values-like or dislike etc. Such values are denoted by the dummy variable. For example, these values can be represented as '1' represents male and '0' represents female. Similarly, '1' represents the liking of taste, and '0' represents the disliking of taste.

3.4 SOURCES OF DATA USED FOR ECONOMIC ANALYSIS

The data used in empirical analysis may be collected by a governmental agency (e.g., the Department of Commerce), an international agency (e.g., the International Monetary Fund [IMF] or the World Bank), a private organization (e.g., the Standard & Poor's Corporation), or an individual. Literally, there are thousands of such agencies collecting data for one purpose or another.

The Internet has literally revolutionized data gathering. If you just "surf the net" with a keyword (e.g., exchange rates), you will be swamped with all kinds of data sources. There are number of frequently visited websites that provide economic and financial data of all sorts. Most of the data can be downloaded without much cost. You may want to bookmark the various websites that might provide you with useful economic data.

The data collected by various agencies may be experimental or nonexperimental. In experimental data, often collected in the natural sciences, the investigator may want to collect data while holding certain factors constant to assess the impact of some factors on a given phenomenon. For instance, in assessing the impact of obesity on blood pressure, the researcher would want to collect data while holding constant the eating, smoking, and drinking habits of the people to minimize the influence of these variables on blood pressure.

In the social sciences, the data that one generally encounters are nonexperimental in nature, that is, not subject to the control of the researcher.

For example, the data on GNP, unemployment, stock prices, etc., are not directly under the control of the investigator. As we shall see, this lack of control often creates special problems for the researcher in pinning down the exact cause or causes affecting a particular situation. Following websites are popular sources of data for economic data:

1. **State of India-CMIE** is a comprehensive compilation of state-level statistics. The data is sourced from each of the 23 major states and 12 minor states or union territories.

2. **Indiastat.com** provides secondary level socio-economic statistical data about India and its states, Region and Sector on more than 35 variables.

3. **CEIC Databases** is delivering a wide range of macroeconomic and industry-specific time series data for India. The database covers over 163,000 time series with historical data from as early as 1951 and offers a wide range of dataset frequencies, from daily to annual.

4. **Open Government Data** Platform India is a platform for supporting Open Data initiative of Government of India. Open Government Data Platform India is also packaged as a product and made available in open source for implementation by countries globally.

5. **Data.gov** provides public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government." Offers numerous free data sets in a searchable database.

6. **UNData:** Do keyword searches to find statistics from the United Nations on many topics including "Agriculture, Crime, Education, Employment, Energy, Environment, Health, HIV/AIDS, Human Development, Industry, Information and Communication Technology, National Accounts, Population, Refugees, Tourism, Trade, as well as the Millennium Development Goals Database

7. **DataHub** provided 8000+ free datasets from the Open Knowledge Foundation. Varied topics. Includes many large datasets from national governments and numerous datasets related to economic development.

8. **Quandloffers** a free platform with hundreds of free data sets from "central banks, exchanges, brokerages, governments, statistical agencies, think-tanks, academics, research firms and more. " You must create a free account on the site to download data. Use Sign Up to create account. EMU does not have access to the premium data on this site, but there are many free data sets.

9. **EconData.net – Guide to Regional Economic Data on the Web:** EconData.Net is designed to help practitioners, researchers, students, and other data users quickly gain access to relevant state and substate socioeconomic data." Categorizes 750+ links to data sources by subject and provider. Subject areas include demographics, employment, occupation, income, output & trade, prices, economic assets, quality of life, industry sectors, and firm listings. Also offers a "free 100-page guide to finding and using economic data to understand your regional economy."

10. **re3data.org - Registry of Research Data Repositories** is a global registry of research data repositories that covers research data repositories from different academic disciplines.

11. **biz/ed Data Sets and Data Skills Materials** provides data and teaching materials for use with students. There is also a Guide To Free Economics And Business Data On The Web. The guide has a British and European focus.

12. **Damodaran Online** offers many downloadable statistics and spreadsheets related to corporate finance and valuation. Includes data on individual firms US &intl, insider trading, historical returns, risk premiums, betas by industry, tax rate comparisons, and more. Caution: although there are past years of data, he has changed methods over time—so the data may not work well for time series analysis.

13. **Federal Reserve Economic Data (FRED)** offers over 148,000 US and international time series from 59 sources.

14. **Global Entrepreneurship Monitor (GEM) project** provides an annual assessment of the entrepreneurial activity, aspirations and attitudes of individuals across a wide range of countries. Can be downloaded to SPSS.

15. **Innovative Data Sources for Economic Analysis** aim is to inform economic researchers and policy makers about new and innovative data sources and analytic tools that have the potential to improve understanding of the dynamics of U.S. economy, specifically as it relates to innovation and entrepreneurship." Scroll down for links to data categories. Some sources described here are not free.

16. **International Macroeconomic Data Set - U.S. Dept of Agriculture Economic Research Service** is useful for projections, the USDA's International Macroeconomic Data Set "provides data from 1969 through 2030 for real (adjusted for inflation) gross domestic product (GDP), population, real exchange rates, and other variables for the 190 countries and 34 regions that are most important for U.S. agricultural trade."

17. **International Monetary Fund Data** IMF time series data for many international economic indicators. Also, data on debt, direct investment, commodities, government finance, exports, exchange rates, etc.

18. **Penn World Table** provides purchasing power parity and national income accounts converted to international prices for 189 countries/ territories for some or all of the years 1950-2010." Provided through the Center for International Comparisons at the University of Pennsylvania.

19. **UK Data service** offers large number of data series — UK, Europe, and international focus.

20. **United Nations Statistical Databases:** "A number of U.N. statistical databases can be accessed for free on this site. Often data can be downloaded. Free sources include data from the Demographic Yearbook System, Joint Oil Data Inititiative, Millennium Indicators Database,

National Accounts Main Aggregates Database (time series 1970-), Social Indicators, population databases, and more. Note additional links to statistical information in the left margin.

21. **World Bank Data Catalog** provides development data, climate change data, GDP data, World Bank finance data, and more.

22. **World Top Incomes Database** - Paris School of Economics provides Distribution of earnings and wealth by country. Can also download data series.

23. **World Resources Institute** is a global research organization that spans more than 50 countries, with offices in Brazil, China, Europe, India, Indonesia, and the United States. Offers a wide range of statistical, graphical, and analytical information related to environmental, social and economic trends.

3.5 THE ACCURACY OF DATA

Although plenty of data are available for economic research, the quality of the data is often not that good. There are several reasons for that.

1. As noted, most social science data are nonexperimental in nature. Therefore, there is the possibility of observational errors, either of omission or commission.

2. Even in experimentally collected data, errors of measurement arise from approximations and roundoffs.

3. In questionnaire-type surveys, the problem of nonresponse can be serious; a researcher is lucky to get a 40 percent response rate to a questionnaire. Analysis based on such a partial response rate may not truly reflect the behaviour of the 60 percent who did not respond, thereby leading to what is known as (sample) selectivity bias. Then there is the further problem that those who do respond to the questionnaire may not answer all the questions, especially questions of a financially sensitive nature, thus leading to additional selectivity bias.

4. The sampling methods used in obtaining the data may vary so widely that it is often difficult to compare the results obtained from the various samples.

5. Economic data are generally available at a highly aggregate level. For example, most macrodata (e.g., GNP, employment, inflation, unemployment) are available for the economy as a whole or at the most for some broad geographical regions. Such highly aggregated data may not tell us much about the individuals or microunits that may be the ultimate object of study.

6. Because of confidentiality, certain data can be published only in highly aggregate form. The IRS, for example, is not allowed by law to disclose data on individual tax returns; it can only release some broad summary data. Therefore, if one wants to find out how much individuals with a certain level of income spent on health care, one cannot do so except at a very highly aggregate level. Such macroanalysis often fails to reveal the dynamics of the behaviour of the microunits. Similarly, the Department of Commerce, which conducts the census of business every 5 years, is not allowed to disclose information on production, employment, energy consumption, research, and development expenditure, etc., at the firm level. It is therefore difficult to study the interfirm differences on these items.

Because of all of these and many other problems, the researcher should always keep in mind that the results of research are only as good as the quality of the data. Therefore, if in given situations researchers find that the results of the research are "unsatisfactory," the cause may be not that they used the wrong model but that the quality of the data was poor. Unfortunately, because of the nonexperimental nature of the data used in most social science studies, researchers very often have no choice but to depend on the available data. But they should always keep in mind that the data used may not be the best and should try not to be too dogmatic about the results obtained from a given study, especially when the quality of the data is suspect.

3.6 SUMMARY

Economic data provide an empirical basis for economic research, whether descriptive or econometric. Data archives are also a key input for assessing the replicability of empirical findings and for use in decision making as to economic policy.

At the level of an economy, many data are organized and compiled according to the methodology of national accounting. Such data include Gross National Product and its components, Gross National Expenditure, Gross National Income in the National Income and Product Accounts, and also the capital stock and national wealth. In these examples data may be stated in nominal or real values, that is, in money or inflation-adjusted terms. Other economic indicators include a variety of alternative measures of output, orders, trade, the labor force, confidence, prices, and financial series (e.g., money and interest rates). At the international level there are many series including international trade, international financial flows, direct investment flows (between countries) and exchange rates.

For time-series data, reported measurements can be hourly (e.g. for stock markets), daily, monthly, quarterly, or annually. Estimates such as averages are often subjected to seasonal adjustment to remove weekly or seasonal-periodicity elements, for example, holiday-period sales and seasonal unemployment.

Within a country the data are usually produced by one or more statistical organizations, e.g., a governmental or quasi-governmental organization and/or the central banks. International statistics are produced by several international bodies and firms, including the International Monetary Fund and the Bank for International Settlements.

Studies in experimental economics may also generate data, rather than using data collected for other purposes. Designed randomized experiments may provide more reliable conclusions than do observational studies.[7] Like epidemiology, economics often studies the behavior of humans over periods too long to allow completely controlled experiments, in which case economists

can use observational studies or quasi-experiments; in these studies, economists collect data which are then analyzed with statistical methods (econometrics).

Many methods can be used to analyse the data. These include, e.g., time-series analysis using multiple regression, Box–Jenkins analysis, and seasonality analysis. Analysis may be univariate (modeling one series) or multivariate (from several series). Econometricians, economic statisticians, and financial analysts formulate models, whether for past relationships or for economic forecasting. These models may include partial equilibrium microeconomics aimed at examining particular parts of an economy or economies, or they may cover a whole economic system, as in general equilibrium theory or in macroeconomics. Economists use these models to understand past events and to forecast future events, e.g., demand, prices and employment. Methods have also been developed for analyzing or correcting results from use of incomplete data and errors in variables.

3.7 GLOSSARY

- **Economic data:** Economic data provide an empirical basis for economic research, whether descriptive or econometric. Data archives are also a key input for assessing the replicability of empirical findings and for use in decision making as to economic policy.
- **Economic statistics:** Economic statistics is a topic in applied statistics and applied economics that concerns the collection, processing, compilation, dissemination, and analysis of economic data. It is closely related to business statistics and econometrics.
- **Macroeconomic management:** The training focuses on such subjects as financial programming and policies, monetary and exchange operations, public finance, financial sector issues, and macroeconomic statistics.
- **International Monetary Fund (IMF):** The International Monetary Fund (IMF) works to achieve sustainable growth and prosperity for all of its 190 member countries. It does so by supporting economic policies that

promote financial stability and monetary cooperation, which are essential to increase productivity, job creation, and economic well-being.

- **General Data Dissemination System (GDDS):** The General Data Dissemination System (GDDS) is a structured process through which IMF member countries commit voluntarily to improving the quality of the data produced and disseminated by their statistical systems over the long run to meet the needs of macroeconomic analysis
- **Time series data:** Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order. Time-stamped is data collected at different points in time. These data points typically consist of successive measurements made from the same source over a time interval and are used to track change over time.
- **GDP:** GDP measures the monetary value of final goods and services—that is, those that are bought by the final user—produced in a country in a given period of time (say a quarter or a year). It counts all of the output generated within the borders of a country.
- **GNP:** Gross National Product (GNP) is the total value of all finished goods and services produced by a country's citizens in a given financial year, irrespective of their location. GNP also measures the output generated by a country's businesses located domestically or abroad.
- **Cross-section data:** Cross-sectional data, or a cross section of a study population, in statistics and econometrics, is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the one point or period of time.
- **Panel data:** Panel data, sometimes referred to as longitudinal data, is data that contains observations about different cross sections across time. Examples of groups that may make up panel data series include countries, firms, individuals, or demographic groups.
- **Dummy variable:** A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study.

In research design, a dummy variable is often used to distinguish different treatment groups.

- **World Bank:** The World Bank is an international development organization owned by 187 countries. Its role is to reduce poverty by lending money to the governments of its poorer members to improve their economies and to improve the standard of living of their people.

3.8 SELF-ASSESSMENT QUESTIONS

Q1. Define International Monetary Fund?

_____

_____

_____

Q2. Define time series data?

_____

_____

_____

Q3. What is GDP?

_____

_____

_____

Q4. Define GNP?

_____

_____

_____

Q5. Explain Cross-section data?

_____

_____

Q6. What is Longitudinal data?

Q7. Explain dummy variable data?

Q8. What is IMF?

Q9. What is economic research?

3.9 LESSON END EXERCISE

Q1. Explain the nature of data used in economic analysis?

Q2. Explain sources of data used for economic analysis?

_____

Q3. Explain the accuracy of data while doing economic analysis for econometrics?

_____

_____

_____

3.10 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.
10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.
11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-I **LESSON NO. 4**

INTRODUCTION OF FINANCIAL ECONOMETRICS

BASICS OF REGRESSION, TWO VARIABLE REGRESSION MODEL - ASSUMPTION, ESTIMATION THROUGH OLS, PROPERTIES OF ESTIMATES, GAUSS-MARKOV THEOREM

STRUCTURE

4.1 INTRODUCTION

The term regression was introduced by Francis Galton. In a famous paper, Galton found that, although there was a tendency for tall parents to have tall children and for short parents to have short children, the average height of children born of parents of a given height tended to move or "regress" toward the average height in the population as a whole. In other words, theheight of the children of unusually tall or unusually short parents tends to move toward the average height of the population. Galton's law of universal regression was confirmed by his friend Karl Pearson, who collected more than a thousand records of heights of members of family groups.He found that the average height of sons of a group of tall fathers was less than their fathers' height and the average height of sons of a group of short fathers was greater than their fathers' height, thus "regressing" tall and short sons alike toward the average height of all men. In the words of Galton, this was "regression to mediocrity."

4.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- Basics of regression,
- assumptions of two variable regression model,
- estimation through OLS,
- Properties of estimates and
- Gauss Markov Theorem.

4.3 THE MODERN INTERPRETATION OF REGRESSION

The modern interpretation of regression is, however, quite different. Broadly speaking, we maysay Regression analysis is concerned with the study of the dependence of one variable, thedependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling)values of the latter.

4.4 OBJECTIVES OF REGRESSION ANALYSIS

1. The key objective behind regression analysis is the statistical dependence of one variable,the dependent variable, on one or more other variables, the explanatory variables.
2. The objective of such analysis is to estimate and/or predict the mean or average value of thedependent variable on the basis of the known or fixed values of the explanatory variables.
3. In practice the success of regression analysis depends on the availability of the appropriatedata.
4. In any research, the researcher should clearly state the sources of the data used in the analysis, their definitions, their methods of collection, and any gaps or omissions in the data as well as any revisions in the data.
5. The data used by the researcher are properly gathered and that the computations and analysisare correct.

4.5 TWO VARIABLE REGRESSION MODEL ANALYSIS

Under single regression model one variable, called the dependent variable is expressed as a linear function of one or more other variable, called explanatory variable. Two variable regression model analysis means a function has only one dependent variable and only one independent variable.

Two variable or bivariate means regression in which the dependent variable (the regressand) is related to a single explanatory variable (the regression).

When mean values depend upon conditioning (variable X) is called conditional expected value. Regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values ofthe explanatory variable (s).

WEEKLY FAMILY INCOME X, $										
$Y_{\downarrow}$ \ $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y \mid X)$	65	77	89	101	113	125	137	149	161	173

To understand this, consider the data given in the below table. The data in the table refer to a total population of 60 families in a hypothetical community & their weekly income (X) and weekly consumption expenditure (Y), both in dollars. The 60 families are divided into 10 income groups (from $80 to $260) and the weekly expenditures of each family in the various groups are as shown in the table. Therefore, we have 10 fixed values of X and the corresponding Y values against each of the X values; and hence there are 10 Y sub populations. There is considerable variation in weekly consumption expenditure in each income group, which can be seen clearly but the general picture that one gets is that, despite the variability of weekly consumption expenditure within each income bracket, on the average, weekly consumption expenditure increases as income increases. To see this clearly, in the given table we have given the mean, or average, weekly consumption expenditure corresponding to each of the 10 levels of income. Thus, corresponding to the weekly income level of $80, the mean consumption expenditure is $65, while corresponding to the income level of $200, it is $137. In all we have 10 mean values for the 10 sub populations of Y. We call these mean values conditional expected values, as they depend on the given values of the (conditioning) variable X. Symbolically, we denote them as E(Y | X), which is read as the expected value of Y given the value of X.

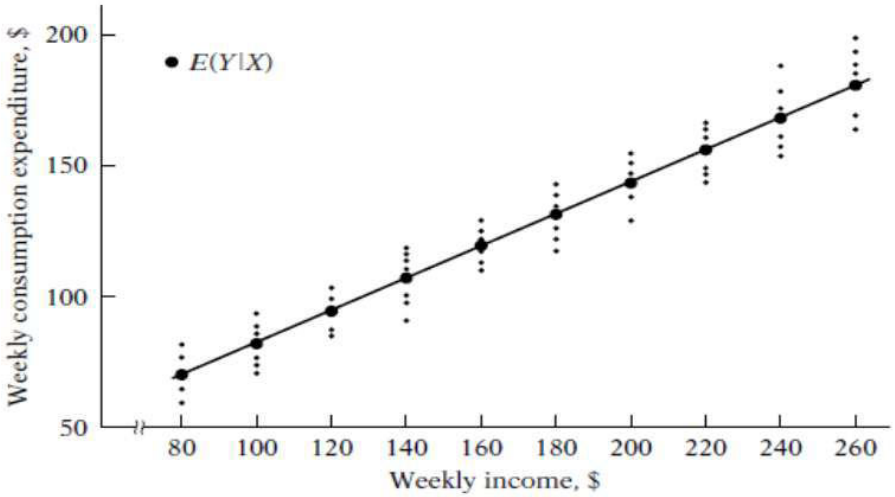


fig.: Conditional distribution of expenditure for various levels of income

It is important to distinguish these conditional expected values from the unconditionalexpected value of weekly consumption expenditure, E(Y). If we add the weekly consumption expenditures for all the 60 families in the population and divide this number by 60, we get the number $121.20 ($7272/60), which is the unconditional mean, or expected, value of weeklyconsumption expenditure, E(Y); it is unconditional in the sense that in arriving at this numberwe have disregarded the income levels of the various families. Obviously, the various conditional expected values of Y given in given table are different from the unconditional expected value of Y of $121.20. When we ask the question, "What is the expected value of weekly consumption expenditure of a family," we get the answer $121.20 (the unconditionalmean). But if we ask the question, "What is the expected value of weekly consumption expenditure of a family whose monthly income is, differently, if we ask the question, "What is the best (mean) prediction of weekly expenditure of families with a weekly income of $140," the answer would be $101.Thus the knowledge of the income level may enable us to better predict the mean value of consumption expenditure

than if we do not have that knowledge.This probably is the essence of regression analysis, as we shall discover throughout this text.

The dark circled points in figure show the conditional mean values of Y against the various X values. If we join these conditional mean values, we obtain what is known as the population regression line (PRL), or more generally, the population regression curve. More simply, it is the regression of Y on X. The adjective "population" comes from the fact that we are dealing in this example with the entire population of 60 families. Of course, in reality a population may have many families.

Geometrically, then, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s). More simply, it is the curve connecting the means of the subpopulations of Y corresponding to the given values of the regressor X. It can be depicted as in figure.

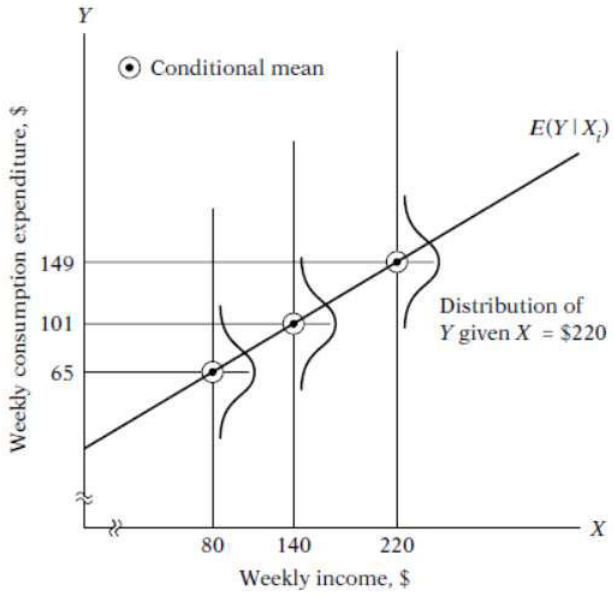


Fig.: Population Regression line.

This figure shows that for each X (i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values.

For simplicity, we are assuming that these Y values are distributed symmetrically around their respective (conditional) mean values. And the regression line (or curve) passes through these (conditional) mean values.

Concept of Population Regression function (PRF) Or Conditional Expectation function

$$\Sigma(Y/X_i)=f(x_i)$$

$f(X_i)$: Some function of the explanatory variable X

$\Sigma(Y/X_i)$: Linear function of X_i

$$\Sigma(Y/X_i)=\beta_1+\beta_2 X_i$$

β_1 & β_2 are unknown but fixed parameters known as the regression coefficients are also known as intercept and slope coefficient. In regression analysis our interest is in estimating the PRFs.

4.6 ESTIMATION THROUGH OLS

Properties of OLS:

1. Our estimation are expressed solely in term of observatory can be easily complete.
2. They are point estimation.
3. Once OLS estimation is obtained from the sample data. The sample regression line can be easily obtained.

$$Y_i=(b_0+b_1x_{1i}+b_2x_{2i})+(u_i)$$

Assumptions of Model:

1. Variable u is real random variable.
2. Homoscedasticity

$$\sum(u_1^2) = \sigma^2$$

3. Normality of u

$$u \sim N(0, \sigma_0^2)$$

4. Non-auto correlation

$$E(u_i u_j) = u \quad i \neq j$$

5. Zero mean of u

$$E(u_i) = 0$$

6. Independence of $E(u_i / x_{1i}) = E(u_i X_{2i}) = 0$
7. No perfect multicollinear X's
8. No error of measurement in the X's.

Estimation through OLS

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}$$

$$Y_i = \hat{Y} + \hat{u} \qquad (Y_i - \hat{u}_i = \hat{Y}_i)$$
$$\hat{u}_i = Y_i - \hat{Y}_i \qquad (\hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i - \hat{u}_i = \hat{Y})$$
$$\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \qquad (\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i)$$

$$\therefore \quad \sum \hat{u}_i^2 = Y_i - \hat{Y}$$

Sq. them we get variation of deviation

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

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

$$\sum \hat{u}_i^2 = \Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\frac{\delta \sum \hat{u}_i^2}{\delta \beta_1} = 2\Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\Sigma Y_i = \Sigma(\hat{\beta}_1 - \hat{\beta}_2 X_i)$$

$$\Sigma Y_i = n\hat{\beta}_1 - \hat{\beta}_2 \Sigma X_i \qquad \text{n= sample size}$$

$$\frac{\delta \sum \hat{u}_i^2}{\delta \beta_2} = 2\Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(X_i) = 0$$

$$X_i \Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

$$\Sigma X_i Y_i = X_i \Sigma(\hat{\beta}_1 - \hat{\beta}_2 X_i)$$

$$\Sigma X_i Y_i = \hat{\beta}_1 X_i - \hat{\beta}_2 X_i^2$$

Note:- We are not taking n β_2 because one variable X1 is already percent. So no need for n, CO_2 they are one & the same.

(LRM) = Classical linear regression Modes) Normal equation Y is dependent upon X. X is independent.)

$$\Sigma Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \Sigma X_i \qquad \rightarrow \quad (2)$$

$$\Sigma X_i Y_i = \hat{\beta}_1 \Sigma X_i + \hat{\beta}_2 \Sigma X_i^2 \rightarrow \quad (3)$$

Dividing equator (2) by n

$$\frac{\sum Y_i}{n} = \frac{n\hat{\beta}_1}{n} + \frac{\hat{\beta}_2 \sum X_i}{n}$$

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$$

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

4.7 PROPERTIES OF ESTIMATES : GAUSS-MARKOV THEOREM

The least-squares estimates possess some ideal or optimum properties; these properties are contained in the well-known Gauss–Markov theorem. To understand this theorem, we need to consider the **best linear unbiasedness property** of an estimator.

BLUE: Best Linear-Unbiased Estimator.

MVUE: Minimum Variance unbiased Estimator.

- If in BLUE, L is not there, because Linearity in co-effects are required not in X &Y. The properties if Least-Square are known as the BLUE.

4.7.1 Properties

1. It is linear i.e. a linear function of a random variable such as the dependent variable Y in the regression model.

2. It is unbiased i.e its average value), $E(\beta_2)$, is = true value of β_2.

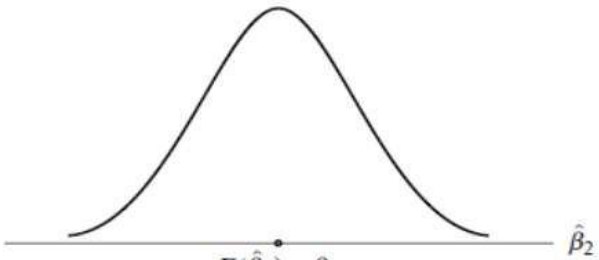
3. Has minimum variance in class of all linear unbiased estimators.

(Note:- An unbiased estimator with the least variance is known as an efficient variable)

4.7.2 Gauss Theorem

Give the assumption of the classical linear regression Model the least squares estimators; in the class of unbiased linear estimator have minimum variance, that is they are BLUE.

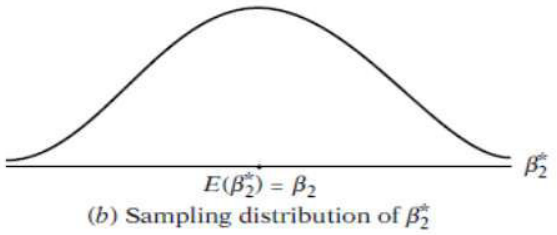
(a) The mean of the β_2 values, $EC(\beta_2)$is equal to the true value of β_2. β_2 is an unbiasedestimator.



(a) Sampling distribution of β_2

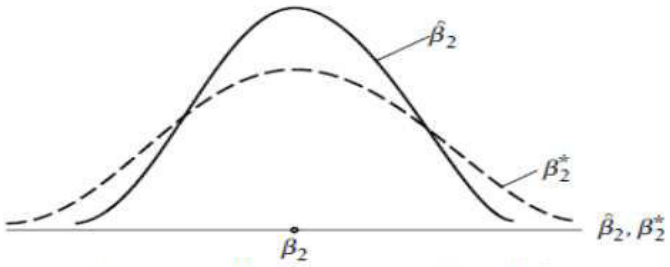
(b)

- Sample distribution of β_2, an alternative estimator of β_2.
- β_2"" & β_2* are linear estimators that is they are linear function of **Y**.
- β_2* like β_2 is unbiased that is, its average or expected value is equal to β_2



(b) Sampling distribution of β_2^*

(c) The variance of β_2 * is larger than the variance of β_2 . One would choose the BLUE estimator.



(c) Sampling distributions of β_2 and β_2^*

G.M. Theorem makes no assumption about the probability distribution of the random variable ui and therefore of $Y_{i.}$.

- As long as the assumption of CLRM are satisfied, the theorem holds.
- If any of the assumption doesn't hold, the theorem is invalid.

4.8 SUMMARY

The key concept underlying regression analysis is the concept of the conditional expectation function (CEF), or population regression function (PRF). Our objective in regression analysis is to find out how the average value of the dependent variable (or regressand) varies with the given value of the explanatory variable (or regressor). This lesson largely deals with linear PRFs, that is, regressions that are linear in the parameters. They may or may not be linear in the regressand or the regressors.For empirical purposes, it is the stochastic PRF that matters. The stochastic disturbance term u_i plays a critical role in estimating the PRF. The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest. Usually, one has a sample of observations from the population. Therefore, one uses the stochastic sample regression function (SRF) to estimate the PRF.In the concluding remarks, we can say that regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the

explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling values of the latter. If we are studying the dependence of a variable on only a single explanatory variable, such as that of consumption expenditure on real income, such a study is known as simple, or two-variable, regression analysis. However, if we are studying the dependence of one variable on more than one explanatory variable, as in the crop-yield, rainfall, temperature, sunshine, and fertilizer examples, it is known as multiple regression analysis. In other words, in two-variable regression there is only one explanatory variable, whereas in multiple regression there is more than one explanatory variable.

4.9 GLOSSARY

- **Regression:** Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
- **Regress:** It means to return to a previous and less advanced or worse state, condition, or way of behaving: She suffered brain damage from the car accident and regressed to the mental age of a five-year-old.
- **Conditional Expectation Function (CEF):** Conditional expectation: the expectation of a random variable X, conditional on the value taken by another random variable Y . If the value of Y affects the value of X (i.e. X and Y are dependent), the conditional expectation of X given the value of Y will be different from the overall expectation of X.
- **Population Regression Function (PRF):** A population regression function hypothesizes a theoretical relationship between a dependent variable and a set of independent or explanatory variables. It is a linear function. The function defines how the conditional expectation of a variable Y responds to the changes in independent variable X.
- **Sample regression function (SRF):** The sample regression function is

an equation that represents the relationship between the Y variable and X variable(s) that is based only on the information in a sample of the population.

4.10 SELF-ASSESSMENT QUESTIONS

Q1. Explain the method of Generalized Least Square (GLS) ?

_____

_____

_____

Q2. What is pure Autocorrelation?

_____

_____

_____

Q3. What is OLS?

_____

_____

_____

Q4. What is FGLS?

_____

_____

_____

4.11 LESSON END EXERCISE

Q1. How to correct pure auto correlation?

_____

_____

_____

Q2. Explain the concept of concept of OLS Versus FGLS?

_____

_____

_____

Q3. Explain the concept of concept of HAC?

_____

_____

_____

Q4. Explain the concept of coexistence of Auto correlation and Heteroscedasticity concept?

_____

_____

_____

4.12 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow, G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.

9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.

10. Maddala, G.S. (1997). Econometrics, McGraw Hill; New York.

11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-I **LESSON NO. 5**

INTRODUCTION OF FINANCIAL ECONOMETRICS

CONCEPT OF R^2; DERIVATION OF R^2, ADJUSTED R^2, DEVIATION FROM CLASSICAL LINEAR, REGRESSION ASSUMPTIONS AND GLS

STRUCTURE

5.1 INTRODUCTION

Coefficient of determination (R^2) is a measure of "Goodness of fit" of the fitted regression line fits the data; that is we shall find out how will the

sample regression line fits the data. The coefficient of determination r^2 (Two variable case) or R^2 (multiple regression) is a sum many measure that tells how the sample regression line will fits the data.

5.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- Concept of Derivation of R^2
- properties of R-Square
- assumptions of Classical Linear Regression Models
- concept of GLS (Generalized least squares)

5.3 DERIVATION OF R^2

R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

R-squared can take any values between 0 to 1. Although the statistical measure provides some useful insights regarding the regression model, the user should not rely only on the measure in the assessment of a statistical model. The figure does not disclose information about the causation relationship between the independent and dependent variables.

In addition, it does not indicate the correctness of the regression model. Therefore, the user should always draw conclusions about the model by analyzing r-squared together with the other variables in a statistical model.

Interpretation of R-Squared

The most common interpretation of r-squared is how well the regression model fits the observed data. For example, an r-squared of 60% reveals that 60% of the data fit the regression model. Generally, a higher r-squared indicates a better fit for the model.

However, it is not always the case that a high r-squared is good for the

regression model. The quality of the statistical measure depends on many factors, such as the nature of the variables employed in the model, the units of measure of the variables, and the applied data transformation. Thus, sometimes, a high r-squared can indicate the problems with the regression model.

A low r-squared figure is generally a bad sign for predictive models. However, in some cases, a good model may show a small value.

There is no universal rule on how to incorporate the statistical measure in assessing a model. The context of the experiment or forecast is extremely important, and, in different scenarios, the insights from the metric can vary.

R^2 shows how well terms (data points) fit a curve or line. Adjusted R^2 also indicates how well terms fit a curve or line but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables,adjustedr-squared will increase.Adjusted R^2 will always be less than or equal to R^2.

You only need R^2 when working with **samples**. In other words, R^2 isn't necessary when you have data from an entire population.

The formula is:

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

where:

- N is the number of points in your data sample.
- K is the number of independent regressors, i.e., the number of variables in your model, excluding the constant.

If you already know R^2 then it's a simple formula to work. However, if you do not already have R^2 then you'll probably not want to calculate this by hand! (If you must, see How to Calculate the Coefficient of Determination).

Meaning of Adjusted R^2: Both R^2 and the adjusted R^2 give you an idea of how many data points fall within the line of the regression equation. However, there is **one main difference** between R^2 and the adjusted R^2: R^2 assumes that

every single variable explains the *variation in the dependent variable*. The adjusted R^2 tells you the percentage of *variation explained by only the independent variables that actually affect the dependent variable*.

5.4 PROPERTIES OF R-SQUARE

1. It is a non-negative quantity.
2. Its limits are $0 \leq r\ 2 \leq 1$.

An r^2 1 means a perfect fit r^2 of 0 means no relation. A quantity closely related to but conceptually very much different from r^2 is the coefficient of correlation, is a measure of the degree of association between two variables. It can be computed from

$$r = \pm\sqrt{r^2}$$

Some of the properties of *r* are as follows:

1. It can be positive or negative, the sign depending on the sign of the term in the numerator of, which measures the sample covariation of two variables.
2. It lies between the limits of "1 and +1; that is, "1 d" r d" 1.
3. It is symmetrical in nature; that is, the coefficient of correlation between X and Y(rXY) is the same as that between Y and X(rYX).
4. It is independent of the origin and scale; that is, if we define X*i=aXi + C and Y*i= bYi + d, where a > 0, b > 0, and c and d are constants, then r between X* and Y* is the same as that between the original variables X and Y.
5. If X and Y are statistically independent the correlation coefficient between them is zero; but if r = 0, it does not mean that two variables are independent. In other words, zero correlation does not necessarily imply independence.

6. It is a measure of linear association or linear dependence only; it has no meaning for describing nonlinear relations.

5.5 ASSUMPTIONS OF CLASSICAL LINEAR REGRESSION MODELS

The following section will give a short introduction about the underlying assumptions of the classical linear regression model (OLS assumptions). Given the Gauss-Markov Theorem we know that the least squares estimator b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators. The Gauss-Markov Theorem is telling us that in a regression model, where the expected value of our error terms is zero, $E(\epsilon_i) = 0$ and variance of the error terms is constant and finite $\sigma^2(\epsilon_i) = \sigma^2 < \infty$ and ϵ_i and ϵ_j are uncorrelated for all *i* and *j* the least squares estimator b_0 and b_1 are unbiased and have minimum variance among all unbiased linear estimators.

Following are the assumptions underlying the Gauss-Markov Theorem in greater depth. In order for a least squares estimator to be BLUE (best linear unbiased estimator) the first four of the following five assumptions have to be satisfied:

Assumption 1: Linear Parameter and correct model specification: Assumption 1 requires that the dependent variable y is a linear combination of the explanatory variables X and the error terms ϵ. Additionally we need the model to be fully specified.

Assumption 2: Full Rank of Matrix X: Assumption 2 requires the matrix of explanatory variables X to have full rank. This means that in case matrix X is a NxK matrix Rank (X) = K.

Assumption 3: Explanatory Variables must be exogenous: Assumption 3 requires data of matrix x to be deterministic or at least stochastically independent of ϵ for all i. In other words, explanatory variables x are not allowed to contain any information on the error terms ϵ, i.e. it must not be possible to explain ϵ through X. Mathematically is assumption 3 expressed as

$$E(\varepsilon_2 \mid X) = 0$$

Assumption 4: Independent and Identically Distributed Error Terms:

Assumption 4 requires error terms to be independent and identically distributed with expected value to be zero and variance to be constant. Mathematically is assumption 4 expressed as

$$\epsilon_i \sim iid(0, \sigma^2)$$

Assumption 5: Normal Distributed Error Terms in Population: Assumption 5 is often listed as a Gauss-Markov assumption and refers to normally distributed error terms ϵ in population. However, assumption 5 is not a Gauss-Markov assumption in that sense that the OLS estimator will still be BLUE even if the assumption is not fulfilled.

5.6 GLS (GENERALIZED LEAST SQUARES)

The generalized least squares (GLS) estimator of the coefficients of a linear regression is a generalization of the ordinary least squares (OLS) estimator. It is used to deal with situations in which the OLS estimator is not BLUE (best linear unbiased estimator) because one of the main assumptions of the Gauss-Markov theorem, namely that of homoskedasticity and absence of serial correlation, is violated. In such situations, provided that the other assumptions of the Gauss-Markov theorem are satisfied, the GLS estimator is BLUE.

The linear regression is

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

where:

- y is an N*1 vector of outputs (N is the sample size);
- X is an N*K matrix of regressors (K is the number of regressors);
- β is the K*1vector of regression coefficients to be estimated;
- ε is an N*1 vector of error terms.

We assume that:

1. X has full rank;

2. $E[\varepsilon \mid X] = 0$

3. $Var[\varepsilon \mid X] = V$, wherre is a N*N symmetric positive definite matrix.

These assumptions are the same made in the Gauss-Markov theorem in order to prove that OLS is BLUE, except for assumption 3.

In the Gauss-Markov theorem, we make the more restrictive assumption that

$$Var[\varepsilon \mid X] = \sigma^2 1$$

where *I* is the N*N identity matrix. The latter assumption means that the errors of the regression are homoscedastic (they all have the same variance) and uncorrelated (their covariances are all equal to zero).

Instead, we now allow for heteroskedasticity (the errors can have different variances) and correlation (the covariances between errors can be different from zero).

The GLS estimator

Since *v* is symmetric and positive definite, there is an invertible matrix $\sum$ such that

$$V = \sum \sum^T$$

If we pre-multiply the regression equation by $\sum^{-1}$

$$\sum^{-1} y = \sum^{-1} X\beta = \sum^{-1} \varepsilon$$

Define

$$\hat{y} = \sum^{-1} y$$

$$X = \sum^{-1} X$$

$$\varepsilon = \sum^{-1} \varepsilon$$

so that the transformed equation can be written as

The following proposition holds

The generalized least squares problem

Remember that the OLS estimator of a linear regression solves the problem

$\hat{\beta}_{OLS} = \arg\min_b (y - Xb)^\top (y - Xb)$ which is called generalized least squares problem.

Proof

The GLS estimator can be shown to solve the problem $\hat{\beta}_{GLS} = \arg\min_b (y - Xb)^\top V^{-1} (y - Xb)$ which is called generalized least squares problem.

Proof

The function to be minimized can be written as

$$\begin{aligned} & (y - Xb)^\top V^{-1} (y - Xb) \\ = & \left[\Sigma^{-1} (y - Xb)\right]^\top \left[\Sigma^{-1} (y - Xb)\right] \end{aligned}$$

It is also a sum of squared residuals, but the original residuals are rescaled by before being squared and summed.

Weighted least squares

When the covariance matrix is diagonal (i.e., the error terms are uncorrelated), the GLS estimator is called weighted least squares estimator (WLS). In this case the function to be minimized becomes

$$(y - Xb)^\top V^{-1} (y - Xb) = \sum_{i=1}^{N} V_{ii}^{-1} (y_i - X_i b)^2$$

where is the -th entry of , is the -th row of, and is the -th diagonal element of . Thus, we are minimizing a weighted sum of the squared residuals, in which each

squared residual is weighted by the reciprocal of its variance. In other words, while estimating, we are giving less weight to the observations for which the linear relationship to be estimated is more noisy, and more weight to those for which it is less noisy.

Feasible generalized least squares

Note that we need to know the covariance matrix in order to actually compute. In practice, we seldom know and we replace it with an estimate. The estimator thus obtained, that is, $\hat{\beta}_{FGLS} = \left(X^{\top}\hat{V}^{-1}X\right)^{-1}X^{\top}\hat{V}^{-1}y$ is called **feasible generalized least squares** estimator.

There is no general method for estimating, although the residuals of a fist-step OLS regression are typically used to compute. How the problem is approached depends on the specific application and on additional assumptions that may be made about the process generating the errors of the regression.

Example A typical situation in which is estimated by running a first-step OLS regression is when the observations are indexed by time. For example, we could assume that is diagonal and estimate its diagonal elements with an exponential moving average $\hat{V}_{i,i} = \alpha\hat{V}_{i-1,i-1} + (1-\alpha)\hat{\varepsilon}_i$ where

In statistics, generalized least squares (GLS) is a technique for estimating the unknown parameters in a linear regression model when there is a certain degree of correlation between the residuals in a regression model. In these cases, ordinary least squares and weighted least squares can be statistically inefficient, or even give misleading inferences. GLS was first described by Alexander Aitken in 1936.

5.7 SUMMARY

The least-squares estimators take on certain properties summarized in

the Gauss–Markov theorem, which states that in the class of linear unbiased estimators, the least-squares estimators have minimum variance. In short, they are BLUE. The precision of OLS estimators is measured by their standard errors. The overall goodness of fit of the regression model is measured by the coefficient of determination, r^2. It tells what proportion of the variation in the dependent variable, or regressand, is explained by the explanatory variable, or regressor. This r^2 lies between 0 and 1; the closer it is to 1, the better is the fit. A concept related to the coefficient of determination is the coefficient of correlation, r. It is a measure of linear association between two variables and it lies between "1 and +1. In last we can say that to find the least-squares estimators, in the class of unbiased linear estimators, have minimum variance, that is, they are BLUE we can use the Gauss–Markov theorem and the coefficient of determination r^2 (two-variable case) or R2 (multiple regression) is a summary measure that tells how well the sample regression line fits the data. The coefficient of determination helps in finding the goodness of fit of the fitted regression line to a set of data; that is, we shall find out how "well" the sample regression line fits the data.

5.8 GLOSSARY

- **R-Squared:** R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).
- **OLS:** In statistics, ordinary least squares (OLS) is a type of linear least squares method for choosing the unknown parameters in a linear regression model (with fixed level-one effects of a linear function of a set of explanatory variables) by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the input dataset and the output of the (linear) function of the independent variable.

5.9 SELF-ASSESSMENT QUESTIONS

Q1. Explain the meaning of derivation of R^2?

_____

_____

_____

Q2. What is the Meaning of Adjusted R_2?

_____

_____

_____

5.10 LESSON END EXERCISE

Q1. Explain the concept of Generalized Least Squares.

_____

_____

_____

Q2. Explain the various properties of R-SQUARE.

_____

_____

_____

Q3. Explain the assumptions of classical linear regression models?

_____

_____

_____

Q4. Explain Gauss-Markov theorem.

_____

_____

5.11 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.
10. Maddala, G.S. (1997). Econometrics, McGraw Hill; New York.
11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-II **LESSON NO. 6**

PROBLEM WITH REGRESSION ANALYSIS

PROBLEM OF HETEROSKEDASTICITY

STRUCTURE

6.1 INTRODUCTION

The statistics involved in data analysis all have one thing in common: trying to find patterns within data. Patterns are essential, not only because they help us process everyday phenomena, but also because they can help us try to make predictions about the future.One of the tools you can use to make predictions about the future is regression modelling.

6.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- the scope of econometrics,
- methodology of econometrics,
- types of econometrics and
- difference between econometrics and statistics

6.3 NATURE OF HETEROSCEDASTICITY

One of the important assumptions of the classical linear regression model is that the variance of each disturbance term ui, conditional on the chosen values of the explanatory variables, is some constant number equal to σ^2. This is the assumption of homoscedasticity, or equal (homo) spread (scedasticity), that is, equal variance. Symbolically,

$$E(ui^2) = \sigma^2 \qquad i = 1, 2, \ldots, n \qquad (1)$$

Figure 6.1 - Homoscedastic disturbances

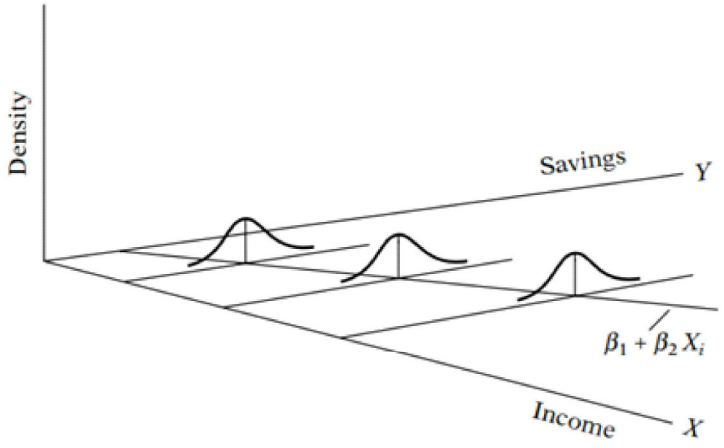
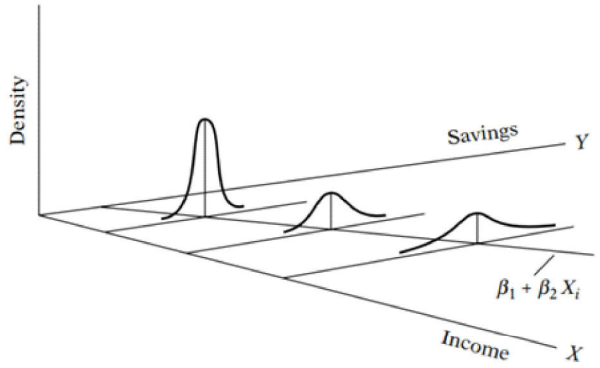


Figure 6.2 Heteroscedastic disturbances



As Figure 6.1 shows, the conditional variance of Yi (which is equal to that of ui), conditional upon the given Xi, remains the same regardless of the values taken by the variable X.

In contrast, consider Figure 6.2, which shows that the conditional variance of Yi increases as X increases. Here, the variances of Yi are not the same. Hence, there is heteroscedasticity. Symbolically,

$$E(ui^2) = \sigma^2 i \qquad (2)$$

Notice the subscript of σ^2, which reminds us that the conditional variances of ui (= conditional variances of Yi) are no longer constant.

To make the difference between homoscedasticity and heteroscedasticity clear, assume that in the two-variable model $Y_i = \beta_1 + \beta_2 Xi + ui$, Y represents savings and X represents income. Figures 6.1 and 6.2 show that as income increases, savings on the average also increase. But in Figure 6.1 the variance of savings remains the same at all levels of income, whereas in Figure 6.2 it increases with income. It seems that in Figure 6.2 the higher-income families on the average save more than the lower-income families, but there is also more variability in their savings.

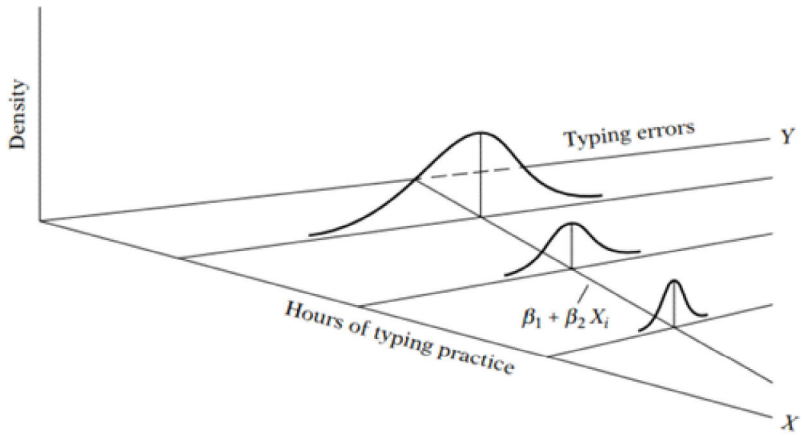
There are several reasons why the variances of ui may be variable, some of which are as follows.

1. Following the error-learning models, as people learn, their errors of behaviour become smaller over time. In this case, σi^2 is expected to decrease. As an example, consider Figure 6.3, which relates the number of typing errors made in a given time period on a test to the hours put in typing practice. As Figure 6.3 shows, as the number of hours of typing practice increases, the average number of typing errors as well as their variances decreases.

2. As incomes grow, people have more discretionary income and hence more scope for choice about the disposition of their income. Hence, σi^2 is likely to increase with income. Thus in the regression of savings on income one is likely to find σi^2 increasing with income (as in Figure 6.2) because people have more choices about their savings behaviour. Similarly,companies with larger profits are generally expected to show greater variability in their dividend policies than companies with lower profits. Also, growth-oriented companies are likely to show more variability in their dividend payout ratio than established companies.

3. As data collecting techniques improve, σi^2 is likely to decrease. Thus, banks that have sophisticated data processing equipment are likely tocommit fewer errors in the monthly or quarterly statements of their customers than banks without such facilities.

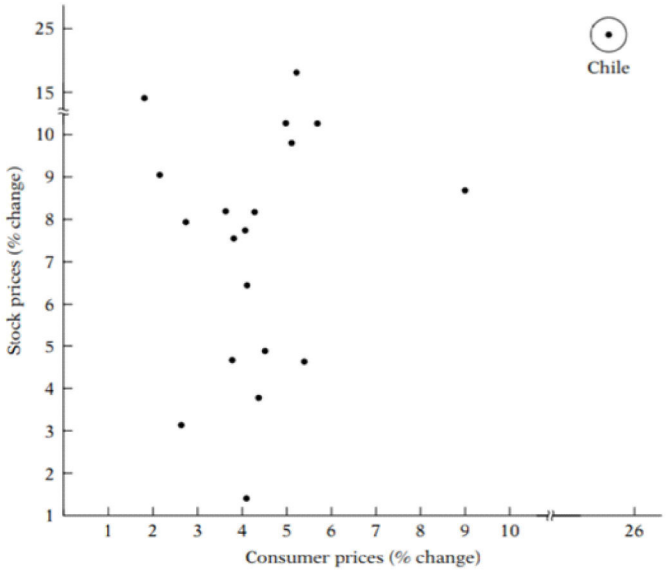
Figure 6.3 Illustration of Heteroscedasticity



4. Heteroscedasticity can also arise as a result of the presence of outliers. An outlying observation, or outlier, is an observation that is much different (either very small or very large) in relation to the observations in the sample. More precisely, an outlier is an observation from a different population to that generating the remaining sample observations. The inclusion or exclusion of such an observation, especially if the sample size is small, can substantially alter the results of regression analysis.

   As an example, consider the scattergram given in Figure 6.4. Based on the exercise 6.1, thisfigure plots percent rate of change of stock prices (Y) and consumer prices (X) for the post–World War II period through 1969 for 20 countries. In this figure the observation on Y and X for Chile can be regarded as an outlier because the given Y and X values are much larger than for the rest of the countries. In situations such as this, it would be hard to maintain the assumption of homoscedasticity. In exercise 6.1, you are asked to find out what happens to the regression results if the observations for Chile are dropped from the analysis.

Figure 6.4 The relationship between stock prices and consumer prices.



5. Another source of heteroscedasticity arises from violating Assumption 9 of CLRM, namely, that the regression model is correctly specified. Very often what looks like heteroscedasticity may be due to the fact that some important variables are omitted from the model. Thus, in the demand function for a commodity, if we do not include the prices of commodities complementary to or competing with the commodity in question (the omitted variable bias), the residuals obtained from the regression may give the distinct impression that the error variance may not be constant. But if the omitted variables are included in the model, that impression may disappear.

   As a concrete example, recall our study of advertising impressions retained (Y) in relation to advertising expenditure (X). (See exercise 6.2.) If you regress Y on X only and observe the residuals from this regression, you will see one pattern, but if you regress Y on X and X2, you will see another pattern, which can be seen clearly from Figure 6.5. We have already seen that X2 belongs in the model.

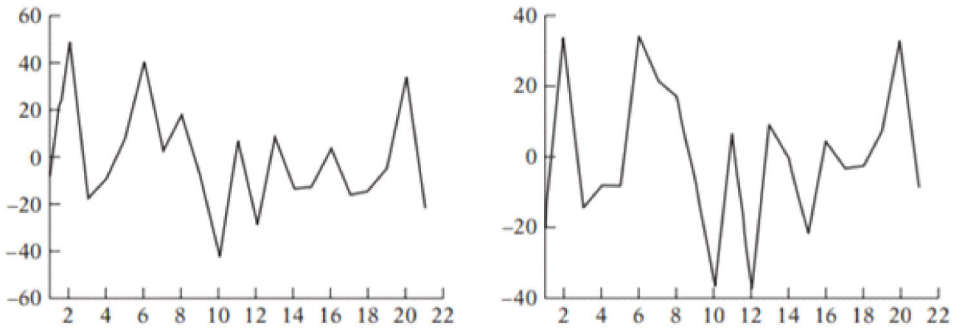
6. Another source of heteroscedasticity is skewness in the distribution of one or more regressors included in the model. Examples are economic variables such as income, wealth, and education. It is well known that the distribution of income and wealth in most societies is uneven, with the bulk of the income and wealth being owned by a few at the top.

7. Other sources of heteroscedasticity: As David Hendry notes, heteroscedasticity can also arise because of (1) incorrect data transformation (e.g., ratio or first difference transformations) and (2) incorrect functional form (e.g., linear versus log–linear models).

Note that the problem of heteroscedasticity is likely to be more common in cross-sectional than in time series data. In cross-sectional data, one usually deals with members of a population at a given point in time, such as individual consumers or their families, firms, industries, or geographical subdivisions such as state, country, city, etc. Moreover, these members may be of different sizes, such as small, medium, or large firms or low, medium, or high income.

In time series data, on the other hand, the variables tend to be of similar orders of magnitude because one generally collects the data for the same entity over a period of time. Examples are GNP, consumption expenditure, savings, or employment in the United States, say, for the period 1950 to 2000.

Figure 6.5 Residuals from the regression of (a) impressions of advertising expenditure and (b) impression on Adexp and Adexp2



As an illustration of heteroscedasticity likely to be encountered in cross-sectional analysis, consider Table 6.1. This table gives data on compensation per employee in 10 nondurable goods manufacturing industries, classified by the employment size of the firm or the establishment for the year 1958. Also given in the table are average productivity figures for nine employment classes.

Although the industries differ in their output composition, Table 6.1 shows clearly that on the average large firms pay more than the small firms.

Table 6.1 Compensation Per Employee ($) in Nondurable Manufacturing Industries According to Employment Size of Establishment, 1958

	Employment size (average number of employees)								
Industry	1–4	5–9	10–19	20–49	50–99	100–249	250–499	500–999	1000–2499
Food and kindred products	2994	3295	3565	3907	4189	4486	4676	4968	5342
Tobacco products	1721	2057	3336	3320	2980	2848	3072	2969	3822
Textile mill products	3600	3657	3674	3437	3340	3334	3225	3163	3168
Apparel and related products	3494	3787	3533	3215	3030	2834	2750	2967	3453
Paper and allied products	3498	3847	3913	4135	4445	4885	5132	5342	5326
Printing and publishing	3611	4206	4695	5083	5301	5269	5182	5395	5552
Chemicals and allied products	3875	4660	4930	5005	5114	5248	5630	5870	5876
Petroleum and coal products	4616	5181	5317	5337	5421	5710	6316	6455	6347
Rubber and plastic products	3538	3984	4014	4287	4221	4539	4721	4905	5481
Leather and leather products	3016	3196	3149	3317	3414	3254	3177	3346	4067
Average compensation	3396	3787	4013	4104	4146	4241	4388	4538	4843
Standard deviation	742.2	851.4	727.8	805.06	929.9	1080.6	1241.2	1307.7	1110.5
Average productivity	9355	8584	7962	8275	8389	9418	9795	10,281	11,750

Source: The Census of Manufacturers, U.S. Department of Commerce, 1958 (computed by author).

As an example, firms employing one to four employees paid on the average about $3396, whereas those employing 1000 to 2499 employees on the average paid about $4843. But notice that there is considerable variability in earning among various employment classes as indicated by the estimated standard deviations of earnings. This can be seen also from Figure 6.6, which plots the standard deviation of compensation and average compensation in each employment class. As can be seen clearly, on average, the standard deviation of compensation increases with the average value of compensation.

Exercise 6.1: Data on percent change per year for stock prices (Y) and consumer prices (X) for a cross section of 20 countries.

a. Plot the data in a scattergram.

b. Regress Y on X and examine the residuals from this regression. What do you observe?

c. Since the data for Chile seem atypical (outlier?), repeat the regression in b, dropping the data on Chile. Now examine the residuals from this regression. What do you observe?

d. If on the basis of the results in b you conclude that there was heteroscedasticity in error variance but on the basis of the results in c you r**everse your conclusion, what general conclusions do you draw?**

Exercise 6.2: Letting Y represent impressions retained and X the advertising expenditure, the following regressions were obtained: Model I: Yˆ i = 22.163 + 0.3631Xi se = (7.089) (0.0971) r2 = 0.424 Model II: Yˆ i = 7.059 + 1.0847Xi ? 0.0040 X2 i se = (9.986) (0.3699) (0.0019) R2 = 0.53 a. Interpret both models. b. Which is a better model? Why? c. Which statistical test(s) would you use to choose between the two models? d. Are there "diminishing returns" to advertising expenditure, that is, after a certain level of advertising expenditure (the saturation level) it does not pay to advertise? Can you find out what that level of expenditure might be? Show the necessary calculations.

6.4 TESTING OF HETEROSCEDASTICITY

The heteroskedasticity-robust standard errors provide a simple method for computing t statistics that are asymptotically t distributed whether or not heteroskedasticity is present. We have also seen that heteroskedasticity-robust F and LM statistics are available. Implementing these tests does not require knowing whether or not heteroskedasticity is present. Nevertheless, there are still some good reasons for having simple tests that can detect its presence. First, the usual t statistics have exact t distributions under the classical linear model assumptions. For this reason, many economists still prefer to see the usual OLS standard errors and test statistics reported, unless there is evidence of heteroskedasticity. Second, if heteroskedasticity is present, the OLS estimator is no longer the best linear unbiased estimator. As we will see in Section 6.3, it is possible to obtain a better estimator than OLS when the form of heteroskedasticity is known. Many tests for heteroskedasticity have been suggested over the years. Some of them, while having the ability to detect heteroskedasticity, do not directly test the assumption that the variance of the

error does not depend upon the independent variables. We will restrict ourselves to more modern tests, which detect the kind of heteroskedasticity that invalidates the usual OLS statistics. This also has the benefit of putting all tests in the same framework.

As usual, we start with the linear model

$$Y_i = \beta\sigma + \beta_{ixi} + \beta_{2x2} + \beta_{ixi} + u, \qquad (1)$$

where Assumptions MLR.1 through MLR.4 are maintained in this section. In particular, we assume that $E(ui|x_1, x_2, \ldots, x_i) = 0$, so that OLS is unbiased and consistent.

We take the null hypothesis to be that Assumption MLR.5 is true:

$$H_0: Var(u\boxtimes_1, x_2, \ldots, x_k) = \sigma^2. \qquad (2)$$

That is, we assume that the ideal assumption of homoskedasticity holds, and we requirethe data to tell us otherwise. If we cannot reject (2) at a sufficiently small significance level, we usually conclude that heteroskedasticity is not a problem. However, remember that we never accept H0; we simply fail to reject it.

Because we are assuming that u has a zero conditional expectation, $Var(u|x) = E(u^2|x)$, and so the null hypothesis of homoskedasticity is equivalent to

$$H_0: E(u^2\boxtimes_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2 . \qquad (3)$$

This shows that, in order to test for violation of the homoskedasticity assumption, we want to test whether u2 is related (in expected value) to one or more of the explanatory variables. If H_0 is false, the expected value of u^2, given the independent variables, can be virtually any function of the x_j. A simple approach is to assume a linear function:

$$u^2 = \delta_0 + \delta_1\boxtimes_1 + \delta_2 x_2 \ldots\ldots + \delta_k x_k + v, \qquad (4)$$

where v is an error term with mean zero given the x_j. Pay close attention to the dependent variable in this equation: it is the square of the error in the original regression equation. The null hypothesis of homoskedasticity is

$$H_0: \delta_1 = \delta_2 = \ldots\ldots \delta_k = 0. \quad (5)$$

Under the null hypothesis, it is often reasonable to assume that the error in (4), v, isindependent of $x_1, x_2, \ldots, x_k$. Then, we know that either the F or LM statistics for the overall significance of the independent variables in explaining u^2 can be used to test (5). Both statistics would have asymptotic justification, even though u^2 cannotbe normally distributed. (For example, if u is normally distributed, then $u^2\sigma^2$ is distributed as X12.) If we could observe the u^2 in the sample, then we could easily compute this statistic by running the OLS regression of u^2 on $x_1, x_2, \ldots, x_k$, using all *n* observations.

As we have emphasized before, we never know the actual errors in the population model, but we do have estimates of them: the OLS residual, $\hat{u}_i$ is an estimate of the error uifor observation i. Thus, we can estimate the equation

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 \ldots\ldots + \delta_k x_k + error \quad (6)$$

and compute the F or LM statistics for the joint significance of $x_1, \ldots, x_k$. It turns out that using the OLS residuals in place of the errors does not affect the large sample distribution of the F or LM statistics, although showing this is pretty complicated.

The F and LM statistics both depend on the R-squared from regression (6); call this R^2 $\hat{u}_i$ to distinguish it from the R-squared in estimating equation (1). Then, the F statistic is

$$F = \frac{R2\ \hat{u}2/k}{(1-R2\ \hat{u}2)/(n-k-1)} \quad (7)$$

where k is the number of regressors in (6); this is the same number of independent variables in (1). Computing (7) by hand is rarely necessary, because most regression packages automatically compute the F statistic for overall

significance of a regression. This F statistic has (approximately) an $F_{k,n-k-1}$ distribution under the null hypothesis of homoskedasticity.

The LM statistic for heteroskedasticity is just the sample size times the R-squared from (6):

$$LM = n \cdot R^2_{\hat{u}^2} \qquad (8)$$

Under the null hypothesis, LM is distributed asymptotically as x_k^2. This is also very easy to obtain after running regression (6).

The LM version of the test is typically called the Breusch-Pagan test for heteroskedasticity (BP test). Breusch and Pagan (1979) suggested a different form of the test that assumes the errors are normally distributed. Koenker (1981) suggested the form of the LM statistic in (8), and it is generally preferred due to its greater applicability. We summarize the steps for testing for heteroskedasticity using the BP test:

The Breusch-Pagan Test for Heteroskedasticity:

1. Estimate the model (1) by OLS, as usual. Obtain the squared OLS residuals, $\hat{u}^2$ (one for each observation).
2. Run the regression in (6). Keep the R-squared from this regression, $R^2_{\hat{u}^2}$.
3. Form either the F statistic or the LM statistic and compute the *p*-value (using the $F_{k,n-k-1}$ distribution in the former case and the xk^2 distribution in the latter case). If the *p*- value is sufficiently small, that is, below the chosen significance level, then we reject the null hypothesis of homoskedasticity. If the BP test results in a small enough *p*-value, some corrective measure should be taken.

The White Test for Heteroskedasticity

Theusual OLS standard errors and test statistics are asymptotically valid, provided all of the Gauss-Markovassumptions hold. It turns out that the homoskedasticity assumption, $Var(u_1|x_1,\ldots, x_k) = \sigma^2$, can be replaced with the

weaker assumption that the squared error, u^2, is uncorrelated with all the independent variables (x_j), the squares of the independent variables (x_j^2), and all the cross products ($x_j x_h$ for $j \neq h$). This observation motivated White (1980) to propose a test for heteroskedasticity that adds the squares and cross products of all the independent variables to equation (6). The test is explicitly intended to test for forms of heteroskedasticity that invalidate the usual OLS standard errors and test statistics. When the model contains k = 3 independent variables, the White test is based on an estimation of

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + error \quad (9)$$

Compared with the Breusch-Pagan test, this equation has six more regressors. The White test for heteroskedasticity is the LM statistic for testing that all of the δ_j in equation (9) are zero, except for the intercept. Thus, nine restrictions are being tested in this case. We can also use an F test of this hypothesis; both tests have asymptotic justification.

With only three independent variables in the original model, equation (9) has nine independent variables. With six independent variables in the original model, the White regression would generally involve 27 regressors (unless some are redundant). This abundance of regressors is a weakness in the pure form of the White test: it uses many degrees of freedom for models with just a moderate number of independent variables.

It is possible to obtain a test that is easier to implement than the White test and more conserving on degrees of freedom. To create the test, recall that the difference between the White and Breusch-Pagan tests is that the former includes the squares and cross products of the independent variables. We can preserve the spirit of the White test while conserving on degrees of freedom by using the OLS fitted values in a test for heteroskedasticity. Remember that the fitted values are defined, for each observation i, by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots\ldots + \beta_k x_{ik}$$

These are just linear functions of the independent variables. If we square the fitted values, we get a particular function of all the squares and cross products of the independent variables. This suggests testing for heteroskedasticity by estimating the equation

$$\square^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + error \qquad (10)$$

where ? stands for the fitted values. It is important not to confuse $\hat{y}$ and y in this equation. We use the fitted values because they are functions of the independent variables (and the estimated parameters); using y in (6.2.10) does not produce a valid test for heteroskedasticity.

We can use the F or LM statistic for the null hypothesis H0: $\delta_1 = 0$, $\delta_2 = 0$ in equation (10). This results in two restrictions in testing the null of homoskedasticity, regardless of the number of independent variables in the original model. Conserving on degrees of freedom in this way is often a good idea, and it also makes the test easy to implement.

6.5 CONSEQUENCES OF HETEROSKEDASTICITY

Consider again the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k + u. \qquad (1)$$

We proved unbiasedness of the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \beta_k$ under the first four Gauss-Markov assumptions, MLR.1 through MLR.4. We also showed that the same four assumptions imply consistency of OLS. The homoskedasticity assumption MLR.5, stated in terms of the error variance as Var $(u|x_1, x_2, \ldots, x_k) = \sigma^2$, played no role in showing whether OLS was unbiased or consistent. It is important to remember that heteroskedasticity does not cause bias or inconsistency in the OLS estimators of the bj, whereas something like omitting an important variable would have this effect.

The interpretation of our goodness-of-fit measures, R^2 and $\bar{R}^2$, is also unaffected by the presence of heteroskedasticity. Theusual R-squared and the adjusted R-squared are different ways of estimating the population R-squared, which is simply $1-\sigma_u^2/\sigma_y^2$, where σ_u^2 is the population error variance and σ_y^2 is the population variance of y. The key point is that because both variances in the population R-squared are unconditional variances, the population R-squared is unaffected by the presence of heteroskedasticity in $Var(u|x_1, \ldots, x_k)$. Further, SSR/n consistently estimates σ_y^2, and SST/n consistently estimates σ_y^2, whether or not $Var(u|x_i, \ldots, x_k)$ is constant. The same is true when we use the degrees of freedom adjustments. Therefore, R^2 and $\bar{R}^2$ are both consistent estimators of the population R-squared whether or not the homoskedasticity assumption holds.

If heteroskedasticity does not cause bias or inconsistency in the OLS estimators, why did we introduce it as one of the Gauss-Markov assumptions ? The estimators of the variances, $Var(\hat{\beta}_j)$, are biased without the homoskedasticity assumption. Since the OLS standard errors are based directly on these variances, they are no longer valid for constructing confidence intervals and *t* statistics. The usual OLS *t* statistics do not have *t* distributions in the presence of heteroskedasticity, and the problem is not resolved by using large sample sizes. We will see this explicitly for the simple regression case in the next section, where we derive the variance of the OLS slope estimator under heteroskedasticity and propose a valid estimator in the presence of heteroskedasticity. Similarly, F statistics are no longer *F* distributed, and the LM statistic no longer has an asymptotic chi-square distribution. In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity.

6.6 REMEDIAL MEASURES OF HETEROSKEDASTICITY

As we have seen, heteroscedasticity does not destroy the unbiasedness andconsistency properties of the OLS estimators, but they are no longer efficient, not even asymptotically (i.e., large sample size). This lack of efficiencymakes

the usual hypothesis-testing procedure of dubious value. Therefore, remedial measures may be called for. There are two approaches to remediation: when σ^2 is known and when σ^2 is not known.

When σ^2 Is Known: The Method of Weighted Least Squares

If σ^2 is known, the most straightforwardmethod of correcting heteroscedasticity is by means of weighted leastsquares, for the estimators thus obtained are BLUE.

When σ^2 Is Not Known

Iftrue σ^2 are known, we can use the WLS method to obtain BLUE estimators. Since the true σ^2 are rarely known, is there a way ofobtaining consistent (in the statistical sense) estimates of the variancesand covariances of OLS estimators even if there is heteroscedasticity? Theanswer is yes.

White's Heteroscedasticity-Consistent Variances and Standard Errors.

White has shown that this estimate can be performed so thatasymptotically valid (i.e., large sample) statistical inferences can be madeabout the true parameter values. Nowadays, several computer packages presentWhite's heteroscedasticity-corrected variances and standard errors alongwith the usual OLS variances and standard errors. Incidentally, White'sheteroscedasticity-corrected standard errors are also known as robuststandard errors.

As the preceding results show, (White's) heteroscedasticity-correctedstandard errors are considerably larger than the OLS standard errors andtherefore the estimated *t* values are much smaller than those obtained byOLS. Based on the latter, both the regressors are statistically significantat the 5 percent level, whereas on the basis of White estimators they are not.However, White's heteroscedasticity-correctedstandard errors can be larger or smaller than the uncorrected standarderrors.

Since White's heteroscedasticity-consistent estimators of the variancesare

now available in established regression packages, it is recommendedthat the reader report them. As Wallace and Silver note:

It is probably a good idea to use the WHITE option [available in regression programs] routinely, perhaps comparing the output with regular OLS output as a check to see whether heteroscedasticity is a serious problem in aparticular set of data.

Plausible Assumptions about Heteroscedasticity Pattern. Apartfrom being a large-sample procedure, one drawback of the White procedureis that the estimators thus obtained may not be so efficient as those obtainedby methods that transform data to reflect specific types of heteroscedasticity. To illustrate this, let us revert to the two-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_1 + u_i$$

We now consider several assumptions about the pattern of heteroscedasticity.

Assumption 1: The error variance is proportional to X2i:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (1)$$

If, as a matter of "speculation," graphical methods, or Park and Glejserapproaches, it is believed that the variance ofui is proportional to thesquare of the explanatory variable X (see Figure 6.6), one may transformthe original model as follows. Divide the original model through by Xi:

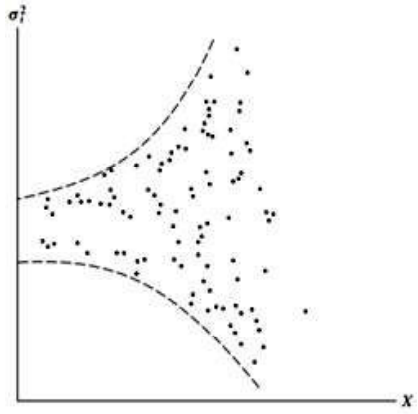
$$\frac{Yi}{Xi} = \frac{\beta 1}{Xi} + \beta 2 + \frac{ui}{Xi}$$

$$= \beta 1 \frac{1}{Xi} + \beta 2 + vi \qquad (2)$$

where vi is the transformed disturbance term, equal to ui/Xi. Now it is easyto verify that

$$E(vi2) = E\left\{\frac{ui}{Xi}\right) = \frac{1}{Xi2}\ E(ii2)\tag{3}$$

Figure 6.6 Error variance proportional to X^2



Hence the variance of vi is now homoscedastic, and one may proceed to apply OLS to the transformed equation (3), regressing Yi/Xi on 1/Xi. Notice that in the transformed regression the intercept term $\beta 2$ is the slope coefficient in the original equation and the slope coefficient $\beta 1$ is the intercept term in the original model. Therefore, to get back to the original model we shall have to multiply the estimated (2) by Xi.

Assumption 2: The error variance is proportional to Xi. The square root transformation:

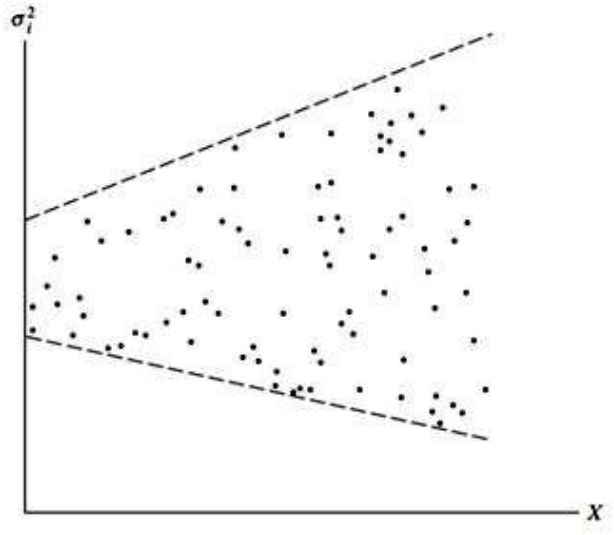
$$E(u_i^2) = \sigma^2 X_i \tag{4}$$

If it is believed that the variance of ui, instead of being proportional to thesquared Xi, is proportional to Xi itself, then the original model can be transformed as follows (see Figure 6.6):

$$\frac{Yi}{\sqrt{Xi}} = \frac{\beta 1}{\sqrt{Xi}} + \beta 2\sqrt{Xi} + \frac{ui}{\sqrt{Xi}}$$
$$= \beta 1 \frac{1}{\sqrt{Xi}} + \beta 2\sqrt{Xi} + vi \tag{4}$$

where $vi = ui/\sqrt{Xi}$ and where $Xi > 0$. .

Figure 6.6 Error variance proportional to X.



Given assumption 2, one can readily verify that $E(v_i^2) = \sigma^2$, a homoscedastic situation. Therefore, one may proceed to apply OLS to (4), regressing *Yi/√Xi* on *1/√Xi* and *√Xi*.

Note an important feature of the transformed model: It has no interceptterm. Therefore, one will have to use the regression-through-the-originmodel to estimate β_1 and β_2. Having run (4), one can get back to theoriginal model simply by multiplying (4) by √Xi.

Assumption 3: The error variance is proportional to the square of the mean value of Y.

$$E(u_i^2) = \sigma^2[E(Y_i)]^2 \qquad (5)$$

Equation (5) postulates that the variance of ui is proportional to thesquare of the expected value of Y. Now

$$E(Y_i) = \beta_1 + \beta_2 X_i$$

Therefore, if we transform the original equation as follows,

$$\frac{Yi}{E(Yi)} = \frac{\beta 1}{E(Yi)} + \beta 2\frac{Xi}{E(Yi)} + \frac{ui}{E(Yi)}$$

$$= \beta 1\left(\frac{1}{E(Yi)}\right) + \beta 2\frac{Xi}{E(Yi)} + vi \qquad (6)$$

where vi=ui/E(Yi), it can be seen that $E(vi^2)= \sigma^2$; that is, the disturbances vi are homoscedastic. Hence, it is regression (6.4.6) that will satisfy the homoscedasticity assumption of the classical linear regression model.

The transformation (6) is, however, inoperational because E(Yi) depends on β_1 and β_2, which are unknown. Of course, we know $\hat{Y}i = \hat{\beta}_1 + \hat{\beta}_2 Xi,$ which is an estimator of E(Yi). Therefore, we may proceed in two steps: First, we run the usual OLS regression, disregarding the heteroscedasticity problem, and obtain $\hat{Y}i$. Then, using the estimated $\hat{Y}i$. We transform our model as follows:

$$\frac{Yi}{\hat{Y}i} = \beta 1\left(\frac{1}{\hat{Y}i}\right) + \beta 2\left(\frac{Xi}{\hat{Y}i}\right) + vi \qquad (7)$$

Where vi=(ui/$\hat{Y}i$) In Step 2, we run the regression (7). Although $\hat{Y}i$ are not exactly E(Yi), they are consistent estimators; that is, as the sample size increases indefinitely, they converge to true E(Yi). Hence, the transformation (7) will perform satisfactorily in practice if the sample size isreasonably large.

Assumption 4: A log transformation such as

$$ln\ Y_i = \beta_1 + \beta_2\ ln\ X_i + u_i \qquad (8)$$

very often reduces heteroscedasticity when compared with the regression $Y_i = \beta_1 + \beta_2 X_i + u_i$

This result arises because log transformation compresses the scalesin which the variables are measured, thereby reducing a tenfold difference between two values to a twofold difference. Thus, the number 80 is10 times the number 8, but ln 80 (= 4.3280) is about twice as large asln 8 (= 2.0794).

An additional advantage of the log transformation is that the slope coefficient β_2 measures the elasticity of Y with respect to X, that is, the

percentage change in Y for a percentage change in X. For example, if Y is consumption and X is income, β_2 in (8) will measure income elasticity, whereas in the original model β_2 measures only the rate of change of mean consumption for a unit change in income. It is one reason why the log models are quite popular in empirical econometrics.

To conclude our discussion of the remedial measures, we reemphasize that all the transformations discussed previously are ad hoc; we areessentially speculating about the nature of σ_i^2. There are some additional problems with the transformations we have considered that should be bornein mind:

1. When we go beyond the two-variable model, we may not know a priori which of the X variables should be chosen for transforming the data.
2. Log transformation as discussed in Assumption 4 is not applicable ifsome of the Y and X values are zero or negative.
3. Then there is the problem of spurious correlation. This term, due toKarl Pearson, refers to the situation where correlation is found to be presentbetween the ratios of variables even though the original variables are uncorrelated or random. Thus, in the model $Yi = \beta_1 + \beta_2 X_i + u_i$, *Y and X* may not be correlated but in the transformed model $Yi / X_i = \beta 1\ (1 / X_i) + \beta 2,\ Yi / X_i$ and $1/ X_i$ are often found to be correlated.
4. When σ_i^2 are not directly known and are estimated from one or moreof the transformations that we have discussed earlier, all our testing procedures using the t tests, F tests, etc., are strictly speaking valid only in largesamples. Therefore, one must be careful in interpreting the results based onthe various transformations in small or finite samples.

6.7 SUMMARY

While heteroscedasticity does not cause bias in the coefficient estimates, it does make them less precise. Lower precision increases the likelihood that

the coefficient estimates are further from the correct population value. Heteroscedasticity tends to produce p-values that are smaller than they should be.The most widely used test for heteroscedasticity is the Breusch-Pagan test. This test uses multiple linear regression, where the outcome variable is the squared residuals. The predictors are the same predictor variable as used in the original model.Heteroskedasticity has serious consequences for the OLS estimator. Although the OLS estimator remains unbiased, the estimated SE is wrong. Because of this, confidence intervals and hypotheses tests cannot be relied on. There are two causes of heteroscedasticity i.e.Omitted variables and mis specified functional form variable. If there are important variables that are not included in the model, this can lead to heteroskedasticity. If the functional form of the model is not correctly specified, this can also lead to heteroskedasticity.Transforming the data is the go-to approach to remove heteroskedasticity. The goal is to stabilize the variance and to bring the distribution closer to the Normal distribution. The log is an effective transformation to do this. Taking the square root or cubic root are two possible alternatives.

6.8 GLOSSARY

- **Regression:** Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
- **Variance:** Variance is a measure of how data points differ from the mean. According to Layman, a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value. Variance means to find the expected difference of deviation from actual value.
- **Homoscedasticity:** Homoscedasticity, or homogeneity of variances, is an assumption of equal or similar variances in different groups being compared. This is an important assumption of parametric statistical tests because they are sensitive to any dissimilarities. Uneven variances in samples result in biased and skewed test results.

- **Heteroscedasticity:** In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant.
- **OLS statistics:** Ordinary Least Squares (OLS) is the best known of the regression techniques. It is also a starting point for all spatial regression analyses. It provides a global model of the variable or process you are trying to understand or predict; it creates a single regression equation to represent that process.
- **Residuals:** Residuals in a statistical or machine learning model are the differences between observed and predicted values of data. They are a diagnostic measure used when assessing the quality of a model. They are also known as errors.

6.9 SELF-ASSESSMENT QUESTIONS

Q1. What you mean by classical linear regression model?

_____

_____

_____

Q2. Differentiate between homoscedasticity and heteroscedasticity.

_____

_____

_____

6.10 LESSON END EXERCISE

Q1. Explain the nature of heteroskedasticity?

_____

_____

_____

Q2. How to test heteroskedasticity?

_____

_____

_____

Q3. What are the consequences of heteroskedasticity?

_____

_____

_____

Q4. Explain various remedial measuresfor the problem of heteroskedasticity.

_____

_____

_____

6.11 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.

10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-II **LESSON NO. 7**

PROBLEM WITH REGRESSION ANALYSIS

PROBLEM OF AUTO CORRELATION

STRUCTURE

7.1 INTRODUCTION

The reader may recall that there are generally three types of data that are available forempirical analysis: (1) cross section, (2) time series, and (3) combination of cross section and time series, also known as pooled data. Incross-section studies, data are often collected based on a randomsample of cross-sectional units, such as households (in a consumption function analysis) orfirms (in an investment study analysis) so that there is no prior reason to believe that theerror term pertaining to one household or firm is correlated with the error term of anotherhousehold or firm. If by chance such a correlation is observed in cross-sectional units, it iscalled spatial autocorrelation, that is, correlation in space rather than over time. However,it is important to remember that, in cross-sectional analysis, the ordering of the data musthave some logic, or economic interest, to make sense of any determination of whether(spatial) autocorrelation is present or not.

The situation, however, is likely to be very different if we are dealing with time series data, for the observations in such data follow a natural ordering over time so that successive observations are likely to exhibit inter correlations, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year. If you observe stock price indexes, such as the Dow Jones or S&P 500, over successive days, it isnot unusual to find that these indexes move up or down for several days in succession. Obviously, in situations like this, the assumption of no auto-, or serial, correlation in the error terms that underlies the CLRM will be violated.

In this chapter we take a critical look at this assumption with a view to describing the nature of auto correlation; test of auto correlation; the oretical and practical consequences of auto correlation; how does one remedy the problem of auto correlation.

7.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- the concept of autocorrelation,

- nature of autocorrelation,
- test used for autocorrelation,
- consequences of autocorrelation,
- remedial measures related to autocorrelation,

7.3 NATURE OF AUTOCORRELATION

1. **Inertia:** Silent feature of most of the time series is inertia or sluggishness. Well known, time series such as GNI price Index.

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

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

where Y = quantity of beef demanded, X2 = price of beef, X3 = consumer income, X4 = price of pork, and t = time.

After Regression:-

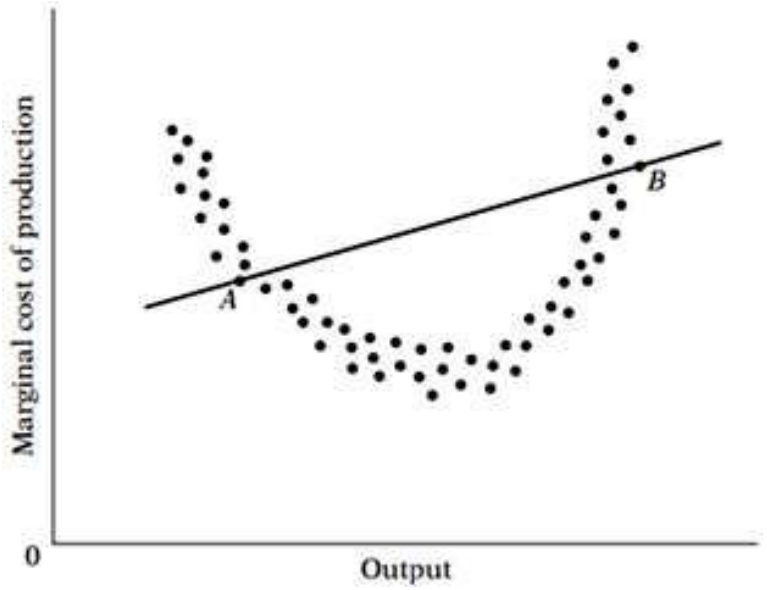
$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + V_t$$

3. **Incorrect Functional Form:**

$$Marginal\ cost_i = \beta_1 + \beta_2\, output_i + \beta_3\, output_i^2 + ui$$

But we get the following model:

Marginal $cost_i = \alpha_1 + \alpha_2\, output_i + v_i$



4. **Cobweb Phenomenon:-** The supply of many agricultural commodities reflects the so-called cobweb Phenomenon. Where supply reacts to price with a lag of one time period because supply decisions takes time implement. Suppl $y_1 = \beta_1 + \beta_2 P_1 1 + u$

5. **Lags**

In a time series regression of consumption expenditure on income, it is not uncommon tofind that the consumption expenditure in the current period depends, among other things,on the consumption expenditure of the previous period. That is,

$$\text{Consumption}_t = \beta_1 + \beta_2 \text{income}_t + \beta_3 \text{consumption}_{t-1} + u_t$$

6. **"Manipulation" of Data**

In empirical analysis, the raw data are often "manipulated". For example, in time series regressions involving quarterly data, such data are usually derived from the monthly databy simply adding three monthly observations and dividing the sum by 3.

7. Data Transformation

As an example of this, consider the following model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \qquad (1)$$

where, say, Y = consumption expenditure and X = income. Since Eq. (7.2.9) holds trueat every time period, it holds true also in the previous time period, (t – 1). So, we can write Eq. (7.2.9) as

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

Y_{t-1},X_{t-1}, and u_{t-1} are known as the lagged values of Y, X, and u, respectively, herelagged by one period. Now if we subtract Eq. (2) from Eq. (1), we obtain

$$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t$$

Where A, known as the first difference operator,

7.4 TEST OF AUTOCORRELATION

7.4.1 Graphical Method:

- Plot any of error
- Error term & there exists non-stationary

Stationary

$Y_t = \rho Y_{t-1} + u_t$

$Y_t = Y_{t-1} + u_t \quad (\rho=1)$

$Y_t - Y_{t-1} = u_t$

Now assume there is lag operation (L)

$(L_{yt} = Y_{t-1})$

$Y_t - LY_t = U_t$

$Y_t (1 - L) = U_t$

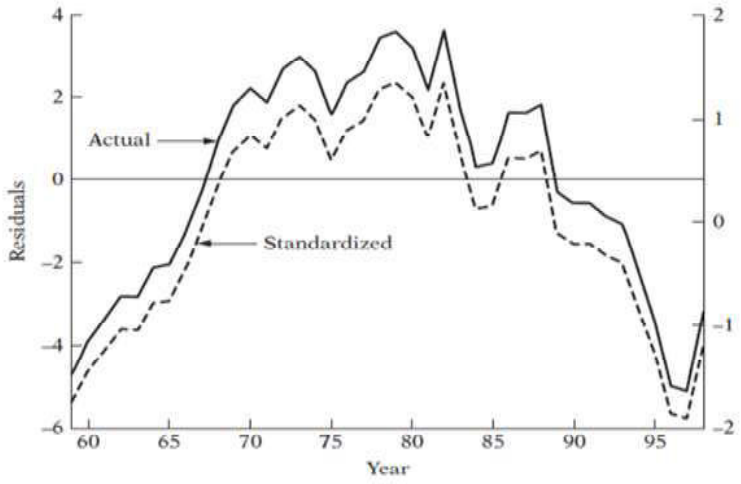
if (1-L) = 0

L = 1

This is known as unit root.

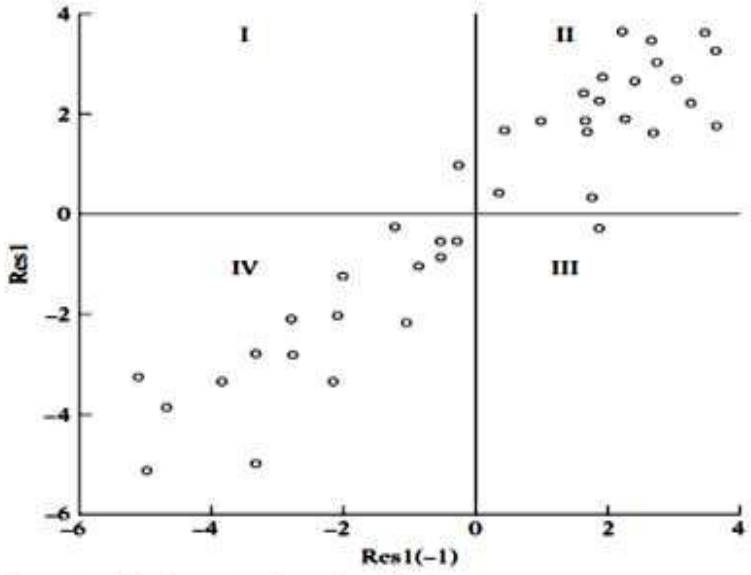
(When root is unit autocorrelation is there) (Nonstationary & unit rest is same) There are various ways of examine the residuals (error)

(a) Time sequence plot



Residuals and standardized residuals from the wages- productivity regression

b) Standardized residual



Current residuals versus lagged residuals.

7.4.2 The Runs Test:

Initially, we have several residuals that are negative, then there is a series of positive residuals, and then there are several residuals that are negative. If these residuals were purely random, could we observe such a pattern? Intuitively, it seems unlikely. This intuition can be checked by the so-called runs test, sometimes also know as the Geary test, a nonparametric test. (------) (++++++++++++++++++++) (----------). This is also a crude method.We now define a run as an uninterrupted sequence of one symbol or attribute, such as + or -. We further define the length of a run as the number of elements in it.

7.4.3 Durbin Watson test:

Also known as Durbin Watson d Test.

One of the good methods as the d statistic is based on the estimated residuals, which are computed in regression analysis

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

This tells where there exists autocorrelation or not

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

$$\simeq 1 + 1 \text{ (by nearly) } - \frac{2\sum \hat{u}_t - \hat{u}_{t-1}}{\sum \hat{u}_t}$$

$$\simeq 2\left(1 - \frac{\sum \hat{u}_t - \hat{u}_{t-1}}{\sum \hat{u}_t}\right)$$

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

$$\begin{cases} 2\,(1-(-1)) = 4 \\ 2\,(1-(1)) = 0 \end{cases}$$

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

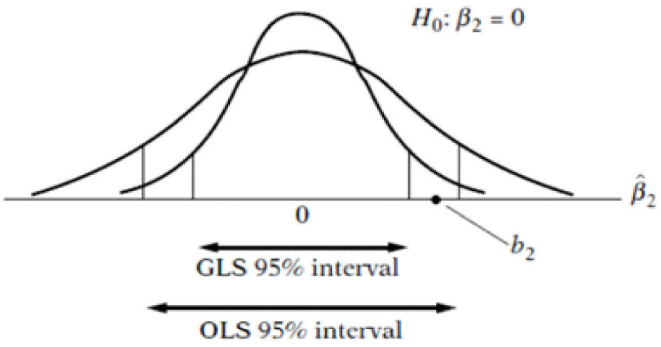
because $p = -1 \leq p \leq 1$

$\rightarrow$ $d \sim 2 \rightarrow$ no autocorrelation

$\rightarrow$ $d \sim 0$ or 4 (closer) there is autocorrelation

7.5 CONSEQUENCES OF AUTOCORRELATION

7.5.1 OLS Estimation allowing for Autocorrelation.



GLS and OLS 95% confidence intervals.

To establish confidence interval to test hypotheses, one should be GLS & not OLS even though the estimators derived from the latter are unbiased & consistent.

7.5.2 Estimation Disregarding Autocorrelation.

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{(n-2)}$$

Unbiased estimate of σ^2 i.e. $\sum(\hat{\sigma}_i^2) = \sigma^2$

$$\sum(\hat{\sigma}^2 - \frac{\sigma^2\{n - [2/(1-\rho)]\} - 2\rho r}{n-2}$$

7.6 REMEDIAL MEASURES FOR AUTOCORRELATION

1. Try to find out if the autocorrelation is pure autocorrelation or not because of the result of the misspecification of the model.

2. Transformation of original model, so that in the transformed model we do not have the problem of (Pure) autocorrelation.

3. In case of large sample, we can Newey-West method to obtain standard error of OLS estimators that are corrected for auto correlation.

4. In some situation we can continue to use the OLS method.

7.7 SUMMARY

Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals.Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals. It measures how the lagged version of the value of a variable is related to the original version of it in a time series. Autocorrelation, as a statistical concept, is also known as serial correlation.This is because autocorrelation can cause problems like: One or more regression coefficients falsely reported as statistically significant. Faux correlations between variables on inferential statistical tests [2]. T-statistics that are too large.The autocorrelation (Box and Jenkins, 1976) function can be used for the following two purposes: To detect non-randomness in data. To identify an appropriate time series model if the data are not random.In astrophysics, autocorrelation is used to study and characterize the spatial distribution of galaxies in the universe and in multi-wavelength observations of low mass X-ray binaries. In panel data, spatial autocorrelation refers to correlation of a variable with itself through space.The existence of autocorrelation in the residuals of a model is a sign that the model may be unsound. Autocorrelation is diagnosed using a correlogram (ACF plot) and can be tested using the Durbin-Watson test.Most statistical tests assume the independence of observations. In other words, the occurrence of one tells nothing about the occurrence of the other. Autocorrelation is problematic for most statistical tests because it refers to the lack of independence between values.

7.8 GLOSSARY

- **Cross sectional data:** Cross-sectional data refer to observations of many different individuals (subjects, objects) at a given time, each observation

belonging to a different individual. A simple example of cross-sectional data is the gross annual income for each of 1000 randomly chosen households in New York City for the year 2000.

- **Time series data:**Time series data, also referred to as time-stamped data, is a sequence of data points indexed in time order. These data points typically consist of successive measurements made from the same source over a fixed time interval and are used to track change over time. .
- **Correlation:** Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).
- **Intercorrelation:** Correlation between the members of a group of variables and especially between independent variables.
- **CLRM:** Classical Linear Regression Model
- **Residuals:** Residuals in a statistical or machine learning model are the differences between observed and predicted values of data. They are a diagnostic measure used when assessing the quality of a model. They are also known as errors.
- **Lags:** This is value of time gap being considered and is called the lag. A lag 1 autocorrelation is the correlation between values that are one time period apart. More generally, a lag k autocorrelation is the correlation between values that are k time periods apart.

7.9 SELF-ASSESSMENT QUESTIONS

Q1. What is cross sectional data ?

_____

_____

_____

Q2. Define time series data.

_____

Q3. What you mean by inertia nature of autocorrelation?

Q4. What you mean by first difference operator?

Q5. Explain graphical method of autocorrelation?

Q6. Explain Run test of autocorrelation.

Q7. Explain Durbin Watson Test.

7.10 LESSON END EXERCISE

Q1. Explain the concept of autocorrelation.

_____

_____

_____

Q2. Explain the nature of autocorrelation.

_____

_____

_____

Q3. How to test autocorrelation?

_____

_____

_____

Q4. Explain the consequences of autocorrelation?

_____

_____

_____

Q5. Explain the remedial measures for autocorrelation?

_____

_____

_____

7.11 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.

5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.

7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.

8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.

9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.

10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-II **LESSON NO. 8**

PROBLEM WITH REGRESSION ANALYSIS

PROBLEM OF MULTICOLLINEARITY

STRUCTURE

8.1 INTRODUCTION

Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables

is high enough, it can cause problems when you fit the model and interpret the results.

8.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- the concept of multicollinearity,
- nature of multicollinearity,
- test used for multicollinearity,
- consequences of multicollinearity,
- remedial measures related to multicollinearity,

8.3 PROBLEM OF MULTICOLLINEARITY

A key goal of regression analysis is to isolate the relationship between each independent variable and the dependent variable. The interpretation of a regression coefficient is that it represents the mean change in the dependent variable for each 1 unit change in an independent variable when you hold all of the other independent variables constant. That last portion is crucial for our discussion about multicollinearity.

The idea is that you can change the value of one independent variable and not the others. However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently because the independent variables tend to change in unison.

8.4 NATURE OF MULTICOLLINEARITY

There are two basic kinds of multicollinearity:

1. **Structural multicollinearity:** This type occurs when we create a model term using other terms. In other words, it's a byproduct of the model

that we specify rather than being present in the data itself. For example, if you square term X to model curvature, clearly there is a correlation between X and X2.

2. **Data multicollinearity:** This type of multicollinearity is present in the data itself rather than being an artifact of our model. Observational experiments are more likely to exhibit this kind of multicollinearity.

8.5 TESTING OF MULTICOLLINEARITY

If you can identify which variables are affected by multicollinearity and the strength of the correlation, you're well on your way to determining whether you need to fix it. Fortunately, there is a very simple test to assess multicollinearity in your regression model. The variance inflation factor (VIF) identifies correlation between independent variables and the strength of that correlation.

Statistical software calculates a VIF for each independent variable. VIFs start at 1 and have no upper limit. A value of 1 indicates that there is no correlation between this independent variable and any others. VIFs between 1 and 5 suggest that there is a moderate correlation, but it is not severe enough to warrant corrective measures. VIFs greater than 5 represent critical levels of multicollinearity where the coefficients are poorly estimated, and the p-values are questionable.

Use VIFs to identify correlations between variables and determine the strength of the relationships. Most statistical software can display VIFs for you. Assessing VIFs is particularly important for observational studies because these studies are more prone to having multicollinearity.

8.6 CONSEQUENCES OF MULTICOLLINEARITY

Multicollinearity causes the following two basic types of problems:

- The coefficient estimates can swing wildly based on which other independent variables are in the model. The coefficients become very sensitive to small changes in the model.

- Multicollinearity reduces the precision of the estimated coefficients, which weakens the statistical power of your regression model. You might not be able to trust the p-values to identify independent variables that are statistically significant.

Imagine you fit a regression model and the coefficient values, and even the signs, change dramatically depending on the specific variables that you include in the model. It's a disconcerting feeling when slightly different models lead to very different conclusions. You don't feel like you know the actual effect of each variable!

Now, throw in the fact that you can't necessarily trust the p-values to select the independent variables to include in the model. This problem makes it difficult both to specify the correct model and to justify the model if many of your p-values are not statistically significant.

As the severity of multicollinearity increases so do these problematic effects. However, these issues affect only those independent variables that are correlated. You can have a model with severe multicollinearity and yet some variables in the model can be completely unaffected.

8.7 REMEDIAL MEASURES FOR THE PROBLEM OF MULTICOLLINEARITY

Multicollinearity makes it hard to interpret your coefficients, and it reduces the power of your model to identify independent variables that are statistically significant. These are serious problems. However, the good news is that you don't always have to find a way to fix multicollinearity.

The need to reduce multicollinearity depends on its severity and your primary goal for your regression model. Keep the following three points in mind:

The severity of the problems increases with the degree of multicollinearity. Therefore, if you have only moderate multicollinearity, you may not need to resolve it.

Multicollinearity affects only the specific independent variables that are correlated. Therefore, if multicollinearity is not present for the independent

variables that you are particularly interested in, you may not need to resolve it. Suppose your model contains the experimental variables of interest and some control variables. If high multicollinearity exists for the control variables but not the experimental variables, then you can interpret the experimental variables without problems.

Multicollinearity affects the coefficients and p-values, but it does not influence the predictions, precision of the predictions, and the goodness-of-fit statistics. If your primary goal is to make predictions, and you don't need to understand the role of each independent variable, you don't need to reduce severe multicollinearity.

8.8 SUMMARY

Multicollinearity might not be severe, it might not affect the variables you're most interested in, or maybe you just need to make predictions. Or perhaps it's just structural multicollinearity that you can get rid of by centering the variables.There are a variety of methods that you can try, but each one has some drawbacks. For this, you can remove some of the highly correlated independent variables; linearly combine the independent variables, such as adding them together; partial least squares regression uses principal component analysis to create a set of uncorrelated components to include in the model.

8.9 GLOSSARY

- **Regression analysis:** Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest. While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.
- **Dependent variable:** A dependent variable is the variable that changes as a result of the independent variable manipulation. It's the outcome you're interested in measuring, and it "depends" on your independent variable. In statistics, dependent variables are also called: Response variables (they respond to a change in another variable)

- **Independent variable:** An independent variable is the variable you manipulate, control, or vary in an experimental study to explore its effects. It's called "independent" because it's not influenced by any other variables in the study. Independent variables are also called: Explanatory variables (they explain an event or outcome).
- **Correlation:** Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.
- **Multicollinearity:** Multicollinearity is a statistical concept where several independent variables in a model are correlated. Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0. Multicollinearity among independent variables will result in less reliable statistical inferences.
- **Structural multicollinearity:** Structural multicollinearity occurs when you use data to create new features. For instance, if you collected data and then used it to perform other calculations and ran a regression on the results, the outcomes will be correlated because they are derived from each other.
- **VIF (Variance Inflation Factor):** 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.

8.10 SELF-ASSESSMENT QUESTIONS

Q1. What is the problem of multicollinearity in statistics?

_____

_____

_____

Q2. Define structural multicollinearity?

_____

_____

_____

Q3. Define data multicollinearity?

_____

_____

_____

Q4. Briefly explain Variance Inflation Factor.

_____

_____

_____

8.11 LESSON END EXERCISE

Q1. Explain the problem of multicollinearity?

_____

_____

_____

Q2. Explain the nature of multicollinearity?

_____

_____

_____

Q3. How to test multicollinearity?

_____

_____

_____

Q4. Explain the consequences of multicollinearity?

_____

_____

_____

Q5. Explain the remedial measures of multicollinearity?

_____

_____

_____

8.12 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow,G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.
10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.
11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-II **LESSON NO. 9**

PROBLEM WITH REGRESSION ANALYSIS

PURE AUTO CORRELATION; OLS VERSUS FGLS AND HAC

STRUCTURE

9.1 INTRODUCTION

Let us study to wages productivity regression. Let suppose the d value is 0.1229 and based on the Durbin-Watson d test we concluded that there was positive correlation in the error term. Since the data underlying regression is time series data, it is quite possible that both wages and productivity exhibit trends. If that is the case, then we need to include the time or trend, t, variable in the model to see the relationship between wages and productivity net of the trends in the two variables.

9.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- concept of Pure Autocorrelation
- the Method of Generalized Least Squares (GLS)
- concept of OLS Versus FGLS and HAC
- concept of HAC
- coexistence of Autocorrelation and Heteroscedasticity

9.3 CORRECTING FOR (PURE) AUTOCORRELATION

9.3.1 The Method of Generalized Least Squares (GLS)

Knowing the consequences of autocorrelation, especially the lack of efficiency of OLS estimators, we may need to remedy the problem. The remedy depends on the knowledge one has about the nature of interdependence among the disturbances, that is, knowledge about the structure of autocorrelation.

As a starter, consider the two-variable regression mode

$$Y_i = \beta 1 + \beta_2 X_i + u_i$$

And assume that the error term follows the AR(1) scheme, namely,

$$\left(u_i - p u_{i-1}\right) = \varepsilon_1 \quad -1 < p < 1$$

Now we consider two cases: (1) p is known and (2) ρ is not known but has to be estimated.

When ρ is known

If the coefficient of first-order autocorrelation is known, the problem of autocorrelation can be easily solved. Hence,

$$Y_{i-1} = \beta 1 + \beta_2 X_{i-1} + u_{t-1}$$

Multiplying by ρ on both sides, we obtain

$$\rho Y_{i-1} = \rho\beta 1 + \rho\beta_2 X_{t-1} + \rho u_{t-1}$$

Subtracting (2) from (1) gives

$$(Y_i - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + \varepsilon_t \qquad 3$$

Where $\varepsilon_t (u_1 - \rho u_{-1})$

We can express (3) as

$$Y_t^2 - \beta_1 + \beta_2 X_t^2 + \varepsilon_t \qquad 4$$

Where $\beta_1 = \beta_1(1-\rho), + Y_t^2 = (Y_t - \rho Y_{t-1}), X_t^2 = (X_t - \rho X_{t-1})$ and $\beta_1 = \beta_2$ 5

Since the error term in (4) satisfies the usual OLS assumptions, we can apply OLS to the transformed variables Y* and X* and obtain estimators with all the optimum properties, namely, BLUE. In effect, running is tantamount to using generalized least squares (GLS) discussed in the previous lesson – recall that GLS is nothing but OLS applied to the transformed model that satisfies the classical assumptions.

Regression (4) is known as the generalized, or quasi, difference equation. It involves regressing Y on X, not in the original form, but in the difference form, which is obtained by subtracting a proportion (=ρ) of the value of a variable in the previous time period from its value in the current time period. In this differencing procedure we lose one observation because the first observation has no antecedent. To avoid this loss of one observation, the first observation on Y and X is transformed as follows. $Y_1\sqrt{1-p^2}$ and $K_1\sqrt{1-p^2}$, This transformation is known as the Prais-Winsten transformation.

9.3.2 OLS Versus FGLS and HAC

The practical problem facing the researcher is this: In the presence of auto-correlation, OLS estimators, although unbiased, consistent, and asymptotically normally distributed, are not efficient. Therefore, the usual inference procedure based on the t, F, and 2 χ tests is no longer appropriate. On the other hand, FGLS (Feasible GLS and EGLS: Estimated GLS) HAC (Heteroscedasticity and autocorrelation estimation) produce estimators that are efficient, but the finite, or small-sample, properties of these estimators are not well documented. This means in small samples the FGLS and HAC might

actually do worse than OLS. As a matter of fact, in a Monte Carlo study Griliches and Rao found that if the sample is relatively small and the coefficient of auto-correlation, ρ, is less than 0.3, OLS is as good or better than FGLS. As a practical matter, then, one may use OLS in small samples in which the estimated rho is, say, less than 0.3. Of course, what is a large and what is a small sample are relative questions, and one has to use some practical judgement. If you have only 15 to 20 observations, the sample may be small, but if you have, say, 50 or more observations, the sample may be reasonably large.

9.4 SUMMARY

Generalized Least Squares (GLS) estimation is a generalization of the Ordinary Least Squares (OLS) estimation technique. GLS is especially suitable for fitting linear models on data sets that exhibit heteroskedasticity (i.e., non-constant variance) and/or auto-correlation.Whereas GLS is more efficient than OLS under heteroscedasticity (also spelled heteroskedasticity) or autocorrelation, this is not true for FGLS.Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).According to Hansen, OLS has the lowest sampling variance among all unbiased estimators, both linear and non-linear. In other words, OLS is BUE—the best unbiased estimator. This is a big deal. The proof that OLS is BLUE, known as the Gauss-Markov theorem, had its initial formulation by Gauss more than 200 years ago.Under the standard assumptions, the OLS estimator in the linear regression model is thus unbiased and efficient. No other linear and unbiased estimator of the regression coefficients exists which leads to a smaller variance. In statistics, ordinary least squares (OLS) is a type of linear least squares method for choosing the unknown parameters in a linear regression model (with fixed level-one effects of a linear function of a set of explanatory variables) by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the input dataset and the output of the (linear) function of the independent variable.

9.5 GLOSSARY

- **GLS:** Generalized Least Squares (GLS) is a method used to estimate the unknown parameters in a linear regression model when there is a certain degree of correlation between the residuals in the regression model.
- **OLS:** Ordinary Least Squares regression (OLS) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).
- **FGLS:** Feasible generalized least squares (FGLS) estimates the coefficients of a multiple linear regression model and their covariance matrix in the presence of non spherical innovations with an unknown covariance matrix.
- **HAC:** Heteroscedasticity and autocorrelation consistent covariance estimators.
- **EGLS:** There are multiple ways to estimate VEC models. A first approach would be to use ordinary least squares, which yields accurate result, but does not allow to estimate the cointegrating relations among the variables. The estimated generalised least squares (EGLS) approach would be an alternative.

9.6 SELF-ASSESSMENT QUESTIONS

Q1. Explain the method of Generalized Least Squares.

_____

_____

_____

Q2. What is pure autocorrelation?

_____

_____

9.7 LESSON END EXERCISE

Q1. Explain the concept of autocorrelation.

Q2. How to correct pure autocorrelation?

Q3. Explain the concept of OLS Versus FGLS and HAC.

Q4. Explain the concept of HAC.

9.8 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow, G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.

5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.

6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.

7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.

8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.

9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.

10. Maddala, G.S.(1997). Econometrics, McGraw Hill; New York.

11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-II **LESSON NO. 10**

PROBLEM WITH REGRESSION ANALYSIS

CO-EXISTENCE OF AUTO CORRELATION AND HETEROSKEDASTICITY

STRUCTURE

10.1 INTRODUCTION

This chapter considers heteroskedasticity and autocorrelation consistent (HAC) estimation of covariance matrices of parameter estimators in linear and nonlinear models. A prime example is the estimation of the covariance matrix of the least squares (LS) estimator in a linear regression model with heteroskedastic, temporally dependent errors of unknown forms. Other examples include covariance matrix estimation of LS estimators of nonlinear regression

models and unit root models and of two and three stage least squares and generalized method of moments estimators of nonlinear simultaneous equations models. HAC estimators have found numerous applications recently in the macro - economic, financial, and international financial literature, e.g., see Campbell and Clarida (1987), Mishkin (1987), and Hardouvelis (1988).

10.2 OBJECTIVES

After studying this lesson, you shall be able to understand:

- concept of coexistence of autocorrelation and heteroscedasticity.
- Concept of dummy variables under heteroscedasticity and autocorrelation

10.3 COEXISTENCE OF AUTOCORRELATION AND HETEROSCEDASTICITY

Heteroscedasticity mainly occurs due to outliers in the data. Besides, Multicollinearity indicates a high correlation between independent variables. Multicollinearity may affect the performance of the regression models. Even some Machine Learning algorithms are too sensitive to Multicollinearity. Heteroscedasticity mainly occurs due to outliers in the data. Besides, Multicollinearity indicates a high correlation between independent variables. Multicollinearity may affect the performance of the regression models. Even some Machine Learning algorithms are too sensitive to Multicollinearity.

10.4 DUMMY VARIABLES UNDER HETEROSCEDASTICITY AND AUTOCORELATION

In previous chapters we discussed the use of dummy variables, but we need to exercise some caution in the use of these variables when we have heteroskedasticity or autocorrelation.

Consider first the case of heteroskedasticity. Suppose that we have the two equations

$$y = \begin{cases} \alpha_1 + \beta_1 x + u_1 & \text{for the first group} \\ \alpha_2 + \beta_2 x + u_2 & \text{for the second group} \end{cases}$$

Let var var $(u_1) = \sigma_1^2$ and var $(u_2) = \sigma_2^2$. When we pool the data, we are implicity assuming that $\sigma_1^2 = \sigma_2^2$. If σ_1^2 and σ_2^2 are widely different, then even if $\square_2$ is not significantly different from $\square_1$ and β_2 is not significantly different from β_1, the coefficients of the dummy variables can turn out to be significant. One can easily demonstrate this by generating data for the two groups imposing $\square_1 = \square_2$ and $\beta_1 = \beta_2$ but $\sigma_1^2 = 16\,\sigma_2^2$ (σ_1^2 being cosen suitably). The reverse situation can also arise; that ts, ignoring heteroskedasticiy can make significant differences appear to be insignificant. Suppose that we take $\square_1 = 2\square_2$ and $\beta_1 = 2\beta_2$. Then by taking $\sigma_1^2\,16\,\sigma_2^2$ (or a multiple around that) we can make the dummy variables appear nonsignificant. The problem is just the same as that of applying tests for stability under heteroskedasticity.

Regarding autocorrelation, suppose that the errors in the equations for the two groups are first-order autoregrssive so that

$$y_t^* = y_t - \rho y_{t-1} \quad \text{for } y_t$$
$$x_t^* = x_t - \rho x_{t-1} \quad \text{for } x_t$$

The question is: What happens to the dummy variables? These variables should not be subject to the autoressive transfromation and care should be taken if the computer program we does this automatically. We can easily derive the appropriate dummy variables in this case.

Consider the case with n_1 observations in the first group and n_2 observations in the second group. We will introduce the time subscript t for each observation later when needed.

Define

$$\alpha_1^* = \alpha_1(1 - \rho)$$
$$\alpha_2^* = \alpha_2(1 - \rho)$$

Then equation can be written as

$$y^* = \alpha_1^* + (\alpha_2^* - \alpha_1^*)D_1 + \beta_1 x^* + (\beta_2 - \beta_1)D_2 + e$$

Where D_2 will be defined as before with x^*_2 in place of x_2 and the random errors et are defined by

$$u_t - \rho u_{t-1} = e_t$$

This equation, however, is all right for the observations in the first group and the last (n_2-1) observations in the second froup. However, the problem is with the first observation in the second group. For this observation, the p-differenced equation turns out to be

$$y_t - \rho y_{t-1} = \alpha_2 - \rho\alpha_1 + (\beta_2 x_t - \rho\beta_1 x_{t-1}) + \acute{e}_t$$

Or

$$y_t^* = \alpha_1^* + \frac{1}{1 - \rho}(\alpha_2^* - \alpha_1^*) + \beta_1 x_t^* + (\beta_2 - \beta_1)x_t + e_t$$

This means that the dummy variables D1 and D2 have to be defined as follows:

$$D_1 = \begin{cases} 0 & \text{for all observations in the first group} \\ \dfrac{1}{1 - \rho} & \text{for the first observation in the second group} \\ 1 & \text{for all the other observations in the second group} \end{cases}$$

$$D_2 = \begin{cases} 0 & \text{for all observations in the first group} \\ x_t & \text{for the first observation in the second group} \\ x_t^* & \text{for all the other observations in the second group} \end{cases}$$

10.5 SUMMARY

Correlation is a measure that is stronger when two things covary more compactly whether they are homoscedastic. Heteroskedasticity is when two things covary in a different pattern of compactness over the ranges of those two things. Autocorrelation refers to the degree of correlation of the same variables between two successive time intervals. It measures how the lagged version of the value of a variable is related to the original version of it in a time series. Autocorrelation, as a statistical concept, is also known as serial correlation.Heteroskedasticity means that the variance of the errors is not constant across observations.The variance of the errors may be a function of explanatory variables. Heteroskedasticity arises if different error terms do not have identical variances, so that the diagonal elements of the covariance matrix are not identical. Autocorrelation almost excessively arises in cases where the data have a time dimension.

10.6 GLOSSARY

- **Coexistence:** The fact of living or existing together at the same time or in the same place:
- **Autocorrelation:** Autocorrelation is a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals.
- **Heteroscedasticity:** In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant.
- **Dummy Variables:** Dummy variables (also known as binary, indicator, dichotomous, discrete, or categorical variables) are a way of incorporating qualitative information into regression analysis. Qualitative data, unlike continuous data, tell us simply whether the individual observation belongs to a particular category.

- **HAC estimators:** The estimator is used to try to overcome autocorrelation (also called serial correlation), and heteroskedasticity in the error terms in the models, often for regressions applied to time series data. The abbreviation "HAC" sometimes used for the estimator, stands for "heteroskedasticity and autocorrelation consistent."
- **Nonlinear regression model:** Nonlinear regression is a statistical technique that helps describe nonlinear relationships in experimental data. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation. Typically machine learning methods are used for non-parametric nonlinear regression.
- **Covariance matrix estimation:** In statistics, sometimes the covariance matrix of a multivariate random variable is not known but has to be estimated. Estimation of covariance matrices then deals with the question of how to approximate the actual covariance matrix on the basis of a sample from the multivariate distribution.
- **Linear and nonlinear models:** In a nonlinear relationship, changes in the output do not change in direct proportion to changes in any of the inputs. While a linear relationship creates a straight line when plotted on a graph, a nonlinear relationship does not create a straight line but instead creates a curve.
- **Multicollinearity:** Multicollinearity is a statistical concept where several independent variables in a model are correlated. Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0. Multicollinearity among independent variables will result in less reliable statistical inferences.

10.7 SELF-ASSESSMENT QUESTIONS

Q1. What is the meaning of Heteroscedasticity?

_____

Q2. What you mean by multicollinearity?

Q3. Define independent variables?

Q4. Briefly define Dummy variable.

Q5. What you mean by Autocorrelation?

Q6. What you mean by HAC?

10.8 LESSON END EXERCISE

Q1. Explain the concept of coexistence of autocorrelation and heteroscedasticity?

_____

_____

_____

Q2. Explain the concept of dummy variables under heteroscedasticity and autocorrelation?

_____

_____

_____

10.9 SUGGESTED READINGS

1. Baltagi, B. Basic Econometrics. Springer, New Delhi.
2. Baltagi, B.H. (1998). Econometrics, Springer, New York.
3. Chow, G.C. (1983). Econometrics, McGraw Hill, New York.
4. Christopher, D. Introduction to Econometrics. Oxford Publishing House, New Delhi.
5. Goldberger, A.S. (1998). Introductory Econometrics, Harvard University Press, Cambridge, Mass.
6. Green, W. (2000). Econometrics, Prentice Hall of India, New Delhi.
7. Gujarati, D.N. (1995). Basic Econometrics. McGraw Hill, New Delhi.
8. Gujarati, D.N., & Sangeetha. Basic Econometrics. Tata McGraw-Hill Publishing Company, New Delhi.
9. Koutsoyiannis, A. (1977). Theory of Econometrics (2nd Esdn.). The Macmillan Press Ltd. London.
10. Maddala, G.S. (1997). Econometrics, McGraw Hill; New York.
11. Ramu, R. Introductory Econometrics with Applications. South-Western College Publishing Company, USA.

****

UNIT-III **LESSON NO. 11**

REGRESSION WITH QUALITATIVE VARIABLE

DUMMY VARIABLES

STRUCTURE

11.1 INTRODUCTION

Dummy variables, also known as indicator variables or binary variables, are often used in statistical modeling and econometrics. They are used to represent categorical data numerically, particularly when the variables are nominal (unordered) rather than ordinal (ordered). Dummy variables are typically created by converting categorical variables into binary variables (0 or 1) to incorporate them into regression models or other statistical analyses.

Here's a brief explanation of how dummy variables work:

1. **Binary Representation:** For a categorical variable with k categories, "1k"1 dummy variables are created. Each dummy variable corresponds to one category and is assigned the value 0 or 1, indicating the absence or presence of that category.

2. **Reference Category:** One category is chosen as the reference category, and its absence is implicitly represented by the dummy variables being 0. The choice of the reference category is arbitrary and does not affect the results of the analysis.

3. **Avoiding Multicollinearity:** Including dummy variables for all categories would introduce perfect multicollinearity, as the values of the dummy variables would be perfectly correlated. By omitting one category as the reference, multicollinearity is avoided.

   Here's a simple example:

   Let's say you have a variable "Color" with three categories: Red, Blue, and Green. You would create two dummy variables, often denoted as Blue$D$Blue and Green$D$Green .

   - Blue=1DBlue =1 if the color is Blue, 0 otherwise.
   - Green=1DGreen =1 if the color is Green, 0 otherwise.

   If both Blue$D$Blue and Green$D$Green are 0, it implies that the color is Red (the reference category).

   In a regression model, for example, the coefficients associated with Blue$D$Blue and Green$D$Green represent the change in the dependent variable when the color is Blue or Green compared to the reference category (Red).

   Dummy variables are crucial in capturing the effects of categorical variables in regression models and other statistical analyses.

   Certainly! Dummy variables are a fundamental concept in statistics, particularly in regression analysis. Here are the basics of dummy variables:

1. **Purpose:**
   - Dummy variables are used to represent categorical data in a numerical form, making it suitable for statistical analysis.
   - They allow for the inclusion of categorical variables in regression models, which typically require numerical input.
2. **Binary Representation:**
   - For a categorical variable with k categories, "1k"1 dummy variables are created.
   - Each dummy variable takes the value of 0 or 1, indicating the absence or presence of a specific category.
3. **Example:**
   - Consider a variable "Gender" with categories: Male and Female.
   - Create a dummy variable Male$D$Male such that Male=1DMale =1 if the observation is male and Male=0DMale =0 otherwise.
   - The absence of Male$D$Male being 1 implies that the observation is female.
4. **Reference Category:**
   - One category is chosen as the reference (baseline) category, and its absence is represented by all dummy variables being 0.
   - The choice of the reference category is arbitrary and doesn't affect the analysis, as it only shifts the interpretation of the coefficients.
5. **Avoiding Multicollinearity:**
   - Including dummy variables for all categories would lead to perfect multicollinearity (perfect correlation) because the sum of dummy variables is constant.
   - By excluding one category as the reference, multicollinearity is avoided, and meaningful interpretations of coefficients become possible.

6. Interpretation of Coefficients:

- In regression models, the coefficients associated with dummy variables represent the change in the dependent variable when moving from the reference category to the represented category.
- For example, if Male*D*Male has a coefficient of 5, it means that being male is associated with an average increase of 5 units in the dependent variable compared to being female (the reference category).

7. Examples of Dummy Variable Coding:

- For a variable with three categories (e.g., A, B, C), two dummy variables (e.g., B*D*B and C*D*C) are created.
- B=1*D*B =1 if the category is B, 0 otherwise.
- C=1*D*C =1 if the category is C, 0 otherwise.

Dummy variables play a crucial role in representing categorical data in a format suitable for statistical analysis, especially in regression models. They allow researchers to incorporate categorical information into their models and draw meaningful conclusions from the results.

11.2 BASICS OF DUMMY VARIABLES

Dummy variables, also known as indicator variables or binary variables, are used in statistical modeling, particularly in regression analysis, to represent categorical data. Here are the basics of dummy variables:

1. Categorical Variables:

- Categorical variables are variables that represent categories or groups. They can be nominal (categories with no inherent order) or ordinal (categories with a meaningful order).

2. Need for Dummy Variables:

- Regression models require numerical input, and many statistical algorithms assume numerical values for variables. Dummy variables are used to represent categorical data numerically, allowing the incorporation of categorical information into regression models.

3. **Binary Representation:**

- Dummy variables take on the values of 0 or 1 to indicate the absence or presence of a particular category. For a categorical variable with "k" categories, you typically create "k-1" dummy variables.

4. **Example:**

- Let's say you have a variable "Color" with three categories: Red, Green, and Blue. You create two dummy variables, say "IsGreen" and "IsBlue." If both "IsGreen" and "IsBlue" are 0 for an observation, it implies that the color is Red.

	Color	IsGreen	IsBlue
	Red	0	0
	Green	1	0
	Blue	0	1

5. **Avoiding Dummy Variable Trap:**

- The dummy variable trap occurs when dummy variables are perfectly correlated, leading to multicollinearity issues in regression analysis. To avoid this, drop one dummy variable, making it a reference category. This reference category is implicitly represented when all other dummy variables are 0.

6. **Use in Regression Models:**

- Dummy variables are included in regression models to account for the effect of categorical variables. Each dummy variable represents a category, and its coefficient in the regression model indicates the average change in the dependent variable associated with that category compared to the reference category.

7. **Interpretation:**

- The coefficients of dummy variables in a regression model represent the average change in the dependent variable associated with a one-unit change in the dummy variable from 0 to 1.

8. Example in Linear Regression:

- In linear regression, if "Y" is the dependent variable and "X" is a dummy variable, the model could be expressed as: $Y = \beta 0 + \beta 1X + \varepsilon$ where $0\beta 0$ is the intercept, $1\beta 1$ is the coefficient associated with the dummy variable, and ε is the error term.

  Dummy variables are a fundamental tool in regression analysis, enabling the incorporation of categorical information into models that require numerical input. Understanding how to properly create and interpret dummy variables is essential for effective statistical modeling.

11.3 IMPORTANCE OF DUMMY VARIABLES

Dummy variables play a crucial role in statistical modeling, particularly in regression analysis. Their importance lies in their ability to represent categorical data, making it possible to include qualitative information in quantitative models. Here are several reasons why dummy variables are important:

1. Handling Categorical Data:

- Many statistical models, including regression analysis, require numerical input. Dummy variables provide a way to represent categorical data, allowing the inclusion of qualitative information in models that operate on numerical values.

2. Incorporating Nominal Data:

- Dummy variables are particularly useful for handling nominal categorical data, where categories have no inherent order. They provide a binary representation for each category, making it possible to include these categories in regression models.

3. Avoiding Ordinal Assumptions:

- When dealing with ordinal categorical data (categories with a meaningful order), dummy variables allow modeling without assuming equal intervals between categories. This flexibility is important for preserving the nature of ordinal data.

4. **Dummy Variable Trap:**

- While it's crucial to avoid the dummy variable trap (perfect multicollinearity), careful handling of dummy variables helps to prevent this issue. By excluding one category as a reference, multicollinearity is mitigated, ensuring stable coefficient estimates.

5. **Interpretability:**

- Dummy variables enhance the interpretability of regression models. Coefficients associated with dummy variables represent the average change in the dependent variable compared to the reference category.

6. **Effect of Categorical Variables:**

- Including dummy variables allows the model to capture the effect of categorical variables on the dependent variable. This is important for understanding how different categories contribute to variations in the outcome.

7. **Statistical Significance Testing:**

- Dummy variables enable statistical hypothesis testing on the significance of different categories. Researchers can assess whether the inclusion of a particular category significantly improves the model fit.

8. **Interaction Effects:**

- Dummy variables can be used to model interaction effects between different categorical variables or between categorical and continuous variables. This allows for a more nuanced understanding of how different factors combine to influence the dependent variable.

9. **Model Flexibility:**

- Dummy variables provide a flexible framework for modeling various types of categorical data, making it possible to include them in a wide range of statistical models beyond linear regression, such as logistic regression or ANOVA.

In summary, dummy variables are essential for incorporating categorical information into quantitative models, offering a versatile and interpretable approach to representing different categories in statistical analyses. Their careful use helps researchers avoid common issues such as the dummy variable trap and ensures meaningful interpretation of regression results.

11.4 TYPES OF DUMMY VARIABLES

Dummy variables are binary variables used to represent categorical data in regression analysis. The number and types of dummy variables depend on the number of categories in the categorical variable. Here are the main types of dummy variables:

1. Binary Dummy Variables:

- These are the most common type of dummy variables and are used for categorical variables with two categories. The dummy variable takes the value of 0 for one category and 1 for the other. For example, in a binary "Gender" variable (Male/Female), you may create a "Is Female" dummy variable.

Gender	Is Female
Male	0
Female	1

2. Multinomial Dummy Variables:

- When dealing with categorical variables with more than two categories, you create multiple binary dummy variables, with one less than the number of categories. The excluded category becomes the reference category. For example, for a "Color" variable with three categories (Red, Green, Blue), you might create "IsGreen" and "IsBlue," with "IsRed" being excluded.

Color	IsGreen	IsBlue
Red	0	0
Green	1	0
Blue	0	1

3. Ordinal Dummy Variables:

- In some cases, you may encounter ordinal categorical variables with a meaningful order. For example, an "Education Level" variable with categories like High School, Bachelor's, Master's, and Ph.D. In this case, you might create ordinal dummy variables where each category represents a level of education.

Education Level	Is Bachelor	Is Master	Is PhD
High School	0	0	0
Bachelor's	1	0	0
Master's	0	1	0
Ph.D.	0	0	1

4. Interaction Dummy Variables:

- Interaction dummy variables are created to represent the interaction between two or more categorical variables. For example, if you have "Region" (North, South, East, West) and "Product Type" (A, B), you might create interaction dummy variables like "Is North Product A," "Is South Product A," and so on.

Region	Product Type	Is North Product A	Is South Product A	Is East Product A	Is West Product A
North	A	1	0	0	0
South	B	0	1	0	0
East	A	0	0	1	0

These are some common types of dummy variables used in regression analysis. The specific structure and naming conventions may vary based on the context and the nature of the categorical variables being represented.

11.5 USES OF DUMMY VARIABLES

Dummy variables are extensively used in statistical modeling, especially in regression analysis, to incorporate categorical data into quantitative models. Here are several common uses of dummy variables:

1. **Regression Analysis:**
   - **Categorical Variables:** Dummy variables are used to represent categorical variables in regression models. This allows the inclusion of qualitative information in models designed for numerical input. For example, if a variable represents different regions (North, South, East, West), dummy variables can be created to code each region.
2. **Market Segmentation:**
   - **Demographic Information:** In market research, dummy variables can be used to represent different demographic groups. This helps in understanding how different groups may respond differently to marketing strategies or product features.
3. **Econometrics:**
   - **Policy Analysis:** Dummy variables are often employed in econometric models to represent policy changes, economic events, or different periods. For example, a dummy variable might indicate the implementation of a new law.
4. **Experimental Design:**
   - **Treatment/Control Groups:** In experimental studies, dummy variables can be used to represent treatment and control groups. This is common in fields such as medicine, where the effect of a treatment is assessed.
5. **Education Research:**

- **Educational Levels:** Dummy variables are used to represent different levels of education in educational research. This allows researchers to examine the impact of education on various outcomes.

6. **Geographical Analysis:**

- **Geographical Regions:** In geographical studies, dummy variables can represent different regions or zones. This is useful for understanding spatial variations in phenomena.

7. **Time Series Analysis:**

- **Seasonal Effects:** In time series analysis, dummy variables can be used to capture seasonal effects, such as holidays or specific months when certain events are likely to occur.

8. **Psychological Studies:**

- **Personality Traits:** Dummy variables can be used to represent different personality traits in psychological studies. Researchers may examine how these traits influence behavior or responses.

9. **Political Science:**

- **Political Affiliation:** In political science research, dummy variables can represent political affiliations (e.g., Democrat, Republican, Independent). This helps analyze the impact of political factors on various outcomes.

10. **Customer Behavior Analysis:**

- **Customer Segmentation:** In marketing, dummy variables can be used to segment customers based on their behaviors, preferences, or demographics. This aids in tailoring marketing strategies to specific customer groups.

11. **Quality Control:**

- **Defective or Non-defective Products:** In manufacturing, dummy variables can be used to indicate whether a product is defective or non-defective. This helps assess the impact of different factors on product quality.

12. Occupational Studies:

- **Occupational Categories:** Dummy variables can represent different occupational categories in studies related to income, job satisfaction, or other work-related outcomes.

  Dummy variables are versatile tools with a wide range of applications in various fields, allowing researchers to include categorical information in their models and analyze the impact of different categories on the dependent variable of interest.

11.6. SUMMARY

Dummy variables, also known as indicator variables or binary variables, are a concept used in statistical modeling and econometrics. They are particularly employed in regression analysis to handle categorical data or factors that can take on two or more distinct values.

Here's a summary of dummy variables:

1. **Purpose:** Dummy variables are introduced to represent categorical data, allowing for the inclusion of qualitative information in regression models.
2. **Binary Encoding:** For a categorical variable with two categories, a single dummy variable is created, taking on values 0 or 1 to indicate the absence or presence of the category.
3. **Multicategory Encoding:** In cases where there are more than two categories, additional dummy variables are generated. For n categories, n-1 dummy variables are created. This avoids multicollinearity issues (perfect correlation) among the variables.
4. **Interpretation:** The coefficient associated with a dummy variable represents the average change in the dependent variable when moving from the reference category (coded as 0) to the category represented by the dummy variable (coded as 1).
5. **Reference Category:** One category is typically chosen as the reference

category, and the dummy variables for other categories are compared to this reference. The choice of the reference category does not affect the model's predictions but can influence the interpretability of coefficients.

6. **Example:** In a regression analysis predicting salary based on education level with three categories (high school, bachelor's, master's), two dummy variables might be created: "Bachelor's degree" and "Master's degree." The reference category, for example, could be "High school," and the coefficients for the dummy variables would indicate the average salary difference for individuals with a bachelor's or master's degree compared to those with only a high school education.

Dummy variables are a valuable tool in statistical modeling, enabling the incorporation of categorical information into regression models and enhancing their ability to explain and predict outcomes.

11.7. GLOSSARY

Dummy Variables: Dummy variables, also known as indicator variables or binary variables, are commonly used in statistical modeling and regression analysis. They are used to represent categorical data with two or more categories in a quantitative manner. In particular, they are often employed when dealing with categorical variables that have no inherent numerical meaning.

Uses of dummy variables in Regression?

Dummy variables are commonly used in regression analysis for handling categorical variables. Categorical variables represent groups or categories, and they may not have a numerical interpretation. Using dummy variables in regression allows you to include categorical data in the model and assess the impact of different categories on the dependent variable.

11.8. SELF ASSESSMENT QUESTIONS

1. What are the dummy variables?

_____

_____

_____

2. Explain the types of Dummy variables?

_____

_____

_____

3. What are the uses of dummy variables?

_____

_____

_____

11.9. SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell. Andrew W. Lo.* A. *Craig MacKinlay,* 1996 "The

Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29, 1996*

10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"
11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"
12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"
13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"
14. Ruud, Paul A., 2000 "An Introduction to Classical Econometric Theory", Oxford, 2000.
15. Wooldridge, J M, 2009 "Introductory Econometrics – A Modern Approach" (4th ed), South-Western, 2009.

****

UNIT-III **LESSON NO. 12**

REGRESSION WITH QUALITATIVE VARIABLE

TESTING STRUCTURAL STABILITY OF REGRESSION MODELS

STRUCTURE

12.1 INTRODUCTION

Testing the structural stability of regression models is an important step to ensure that the relationships between the independent variables and the dependent variable remain consistent over time or across different subgroups. Structural stability tests help assess whether the model's parameters are stable or if there are significant changes in the relationships between variables.

Here are some common methods to test the structural stability of regression models:

1. **Chow Test:**
   - The Chow test is commonly used to detect structural breaks in regression models. It involves estimating the model parameters for different subgroups or time periods and comparing the sum of squared residuals from the combined model with the sum of squared residuals from separate models for each subgroup or time period.
   - A significant difference suggests structural instability.
2. **CUSUM Test (Cumulative Sum of Residuals):**
   - The CUSUM test involves plotting the cumulative sum of residuals over time. A sudden change or a significant deviation from a constant trend may indicate a structural break.
   - The CUSUM test is often used for time series data.
3. **Recursive Residuals Test:**
   - This test involves estimating the model parameters recursively over time. The recursive residuals are then examined for stability.
   - Significant changes in the recursive residuals may indicate structural instability.
4. **Variance Inflation Factor (VIF):**
   - VIF measures the extent to which the variance of the estimated regression coefficients increases when your predictors are correlated. High VIF values may indicate multicollinearity, which can affect the stability of the model.
5. **Heteroscedasticity Tests:**
   - Heteroscedasticity, or non-constant variance of residuals, can be a sign of structural instability. Tests like the Breusch-Pagan test or the White test can help detect heteroscedasticity.

6. **Interaction Effects:**

- Include interaction terms in the model to account for potential changes in the relationships between variables across different conditions or time periods.

7. **Outlier Analysis:**

- Identify and analyze outliers, as they can influence the stability of the regression model.

8. **Cross-Validation:**

- Use cross-validation techniques to assess how well the model generalizes to new data. Changes in performance metrics across different subsets of the data may indicate structural instability.

It's essential to consider the nature of your data and the specific context of your regression analysis when choosing a method for testing structural stability. Additionally, no single test can guarantee the absence of structural instability; a combination of methods is often more informative.

12.2. OBJECTIVES

Testing the structural stability of regression models is essential to ensure the reliability and validity of the model's predictions over time or across different subsets of data. The main objectives of testing structural stability in regression models include:

1. **Assessing Model Robustness:**

- Ensure that the regression model remains robust and effective when applied to different time periods or subsets of the data.
- Verify that the model's performance is not overly influenced by specific observations or time periods.

2. **Temporal Stability:**

- Evaluate if the relationships captured by the model are consistent over time.

- Identify any temporal patterns or changes in the relationship between the independent and dependent variables.

3. Cross-Sectional Stability:

- Determine if the model's structure holds across different subsets or groups within the data (e.g., different geographical regions, demographic groups, or industries).
- Verify that the model is not biased or overly tailored to a specific subgroup.

4. Identifying Structural Breaks:

- Detect any significant changes or breaks in the relationships between variables.
- Recognize structural shifts in the data that may impact the model's predictive accuracy.

5. Model Validity:

- Ensure that the model's assumptions are valid across various conditions and time frames.
- Assess the model's ability to generalize to new data or different contexts.

6. Prediction Accuracy:

- Maintain consistent prediction accuracy over time and across different subsets of data.
- Identify and address any deterioration in the model's predictive performance due to structural changes.

7. Adaptation to Changes:

- Understand how well the model adapts to changes in the underlying data-generating process.
- Determine if the model needs updates or adjustments to remain relevant in changing conditions.

8. Statistical Significance:

- Confirm that the coefficients of the regression model remain statistically significant and meaningful across different segments of the data.
- Avoid relying on coefficients that may be sensitive to specific observations or periods.

9. Model Transparency and Interpretability:

- Ensure that the model remains interpretable and transparent, allowing stakeholders to understand the relationships between variables.

10. Decision-Making Confidence:

- Provide stakeholders with confidence in using the regression model for decision-making by demonstrating its stability under various conditions.

By addressing these objectives, one can enhance the reliability and applicability of regression models in different contexts and over time, making them more valuable for decision-making and analysis.

12.3 IMPORTANCE OF TESTING STRUCTURAL STABILITY OF REGRESSION MODELS

Testing the structural stability of regression models is crucial for several reasons, and it plays a significant role in ensuring the validity and reliability of the model's results. Here are some important reasons highlighting the significance of testing structural stability:

1. Model Reliability Over Time:

- Regression models are often built on historical data, assuming that the relationships between variables remain stable over time. Testing for structural stability helps ensure that the model remains reliable as conditions and relationships between variables may change.

2. Policy and Economic Changes:

- Economic and policy changes can have a profound impact on relationships between variables. Testing for structural stability is

essential when economic or policy shifts occur, as it helps assess whether the existing model is still valid under the new conditions.

3. **Forecasting Accuracy:**

- Structural stability testing is crucial for forecasting accuracy. If a model is built on historical data but the structure of the relationships changes, predictions based on that model may be inaccurate. Ensuring stability enhances the model's ability to provide reliable forecasts.

4. **Avoiding Model Obsolescence:**

- Over time, external factors or underlying relationships may evolve, rendering the existing model obsolete. Regular structural stability testing helps identify when a model needs updating or re-specification to maintain its relevance.

5. **Financial Risk Management:**

- In finance, where regression models are often used for risk management and portfolio optimization, structural stability testing is crucial. Changes in market conditions, monetary policies, or other external factors can impact the relationships between financial variables.

6. **Policy Evaluation:**

- In policy evaluation studies, testing for structural stability is important to assess whether the impact of a policy remains consistent over time. This is crucial for policymakers to make informed decisions and adjustments.

7. **Model Interpretability:**

- Stability testing contributes to the interpretability of regression models. If a model is stable over time, it provides more confidence in the interpretation of coefficients and relationships between variables.

8. **Avoiding Misleading Conclusions:**

- Failure to test for structural stability may lead to misleading conclusions. If a model assumes stability when the underlying relationships have

changed, the results may not accurately reflect the current state of affairs, potentially leading to incorrect decisions.

9. Regulatory Compliance:

- In regulated industries or when models are used for compliance purposes, demonstrating the stability of the model structure is often a requirement. It helps ensure that the model adheres to regulatory guidelines over time.

10. Scientific Rigor:

- For academic research and scientific studies, testing structural stability enhances the rigor of the analysis. It demonstrates the researcher's commitment to validating the model under changing conditions, contributing to the robustness of the research findings.

In summary, testing the structural stability of regression models is essential for maintaining the relevance and accuracy of the model over time, especially in dynamic environments where relationships between variables may evolve. It is a critical step in ensuring the validity and reliability of the model's results for decision-making and forecasting.

12.4 USES OF TESTING STRUCTURAL STABILITY OF REGRESSION MODELS

Testing the structural stability of regression models is crucial in various fields and scenarios to ensure the reliability and validity of the models over time. Here are some specific uses and applications of testing structural stability in regression models:

1. Economic Forecasting:

- In macroeconomic models, testing structural stability is essential to ensure that relationships between economic variables remain consistent over time. Changes in economic policies, global events, or other factors can affect the stability of economic models.

2. **Financial Modeling:**

- In finance, structural stability testing is crucial for models used in risk management, portfolio optimization, and asset pricing. Fluctuations in market conditions, interest rates, or financial regulations can impact the stability of financial models.

3. **Policy Evaluation:**

- When assessing the impact of government policies or interventions, testing for structural stability helps determine whether the effects of the policy remain constant or if there are changes over time. This is critical for policymakers to make informed decisions.

4. **Business Forecasting:**

- In business and industry, regression models are often used for forecasting sales, demand, or other key performance indicators. Testing structural stability ensures the accuracy of these forecasts, especially when there are shifts in market dynamics or changes in consumer behavior.

5. **Marketing Research:**

- Marketing models that analyze the impact of advertising, promotions, or market segmentation may require structural stability testing. Changes in consumer preferences, market competition, or advertising effectiveness can influence the stability of these models.

6. **Environmental and Climate Modeling:**

- Regression models used in environmental studies or climate modeling may need structural stability testing to account for changes in environmental conditions, regulations, or technological advancements that can affect the relationships between variables.

7. **Healthcare Research:**

- Regression models in healthcare, such as those predicting patient outcomes or assessing the effectiveness of medical treatments, may require structural stability testing. Changes in healthcare policies,

advancements in medical technology, or shifts in patient demographics can impact model stability.

8. **Social Sciences Research:**

- In social sciences, models predicting human behavior, social trends, or public opinion may benefit from structural stability testing. Changes in societal norms, political climates, or cultural shifts can influence the stability of these models.

9. **Energy and Resource Management:**

- Models in energy and resource management, such as those predicting energy consumption or resource utilization, may need structural stability testing. Changes in regulations, technological innovations, or economic factors can affect these models.

10. **Quality Control in Manufacturing:**

- Regression models used in manufacturing for quality control may require structural stability testing. Changes in production processes, materials, or equipment can impact the relationships between variables.

11. **Sports Analytics:**

- Regression models in sports analytics that predict player performance, team outcomes, or fan engagement may benefit from structural stability testing. Changes in team composition, coaching strategies, or rule modifications can influence model stability.

Testing for structural stability in these contexts ensures that regression models remain relevant, accurate, and reliable in the face of changing conditions, thus enhancing the utility of these models for decision-making and analysis.

12.5 TYPES OF TESTING STRUCTURAL STABILITY OF REGRESSION MODELS

Testing the structural stability of regression models is important to ensure that the relationships between variables remain consistent over time or across different subsets of data. Structural stability testing helps assess whether the

model's performance and parameter estimates are robust and reliable. Here are some common approaches for testing the structural stability of regression models:

1. **Chow Test:**
   - The Chow test is a statistical test that compares the regression parameters from different subsets of the data.
   - It involves splitting the data into two or more groups, estimating separate regression models for each group, and then comparing the sum of squared residuals of the separate models with the sum of squared residuals of a combined model.
   - A significant difference in the sums of squared residuals may indicate structural instability.
2. **CUSUM Test:**
   - The Cumulative Sum (CUSUM) test is used to detect changes in the coefficients of a regression model over time.
   - It involves calculating the cumulative sum of the differences between the observed and expected values of the dependent variable.
   - Significant departures from zero in the CUSUM plot may indicate structural breaks in the model.
3. **Recursive Residuals Test:**
   - In the recursive residuals test, the model is estimated over a moving window of observations, and the residuals are examined over time.
   - Changes in the pattern of residuals may indicate structural instability.
   - It's important to correct for multiple testing to avoid false positives.
4. **Cointegration Test:**
   - Cointegration tests are often used when dealing with time series data.
   - These tests check whether there is a long-term relationship between variables, and changes in this relationship can indicate structural instability.

5. Bayesian Methods:

- Bayesian methods can be employed to model uncertainty and incorporate prior beliefs about structural stability.
- Bayesian structural stability tests may involve comparing posterior distributions of parameters across different time periods or subsets of data.

6. Rolling Regression:

- The rolling regression involves estimating the model over a rolling window of observations.
- Changes in the estimated coefficients or model fit over time can indicate structural instability.

7. Outlier Analysis:

- Identifying and analyzing outliers in the data can help assess their impact on the model's stability.
- Outliers may indicate structural breaks or changes in the underlying data generation process.

8. Bootstrap Methods:

- Bootstrap resampling techniques can be used to estimate the uncertainty around parameter estimates.
- Comparing parameter estimates and their confidence intervals across different bootstrap samples can provide insights into structural stability.

It's crucial to note that the choice of a specific test depends on the nature of the data and the suspected sources of structural instability. Additionally, structural stability testing should be accompanied by a thorough understanding of the context and domain knowledge.

12.6 SUMMARY

Testing structural stability in regression models is essential to ensure the reliability and robustness of the relationships captured by the model. Here's

a summary of key points regarding the testing of structural stability in regression models:

1. **Objective:**
   - The primary goal is to assess whether the relationships between variables remain consistent over time or across different subsets of data.
2. **Common Tests:**
   - Various statistical tests are employed to detect structural instability, including the Chow test, CUSUM test, recursive residuals test, cointegration test, Bayesian methods, rolling regression, outlier analysis, and bootstrap methods.
3. **Chow Test:**
   - Compares regression parameters from different subsets of data and identifies significant differences in the sum of squared residuals, indicating structural instability.
4. **CUSUM Test:**
   - Detects changes in regression coefficients over time by analyzing the cumulative sum of differences between observed and expected values.
5. **Recursive Residuals Test:**
   - Estimates the model over moving windows of data, observing changes in residual patterns as potential indicators of structural instability.
6. **Cointegration Test:**
   - Especially relevant for time series data, cointegration tests assess long-term relationships between variables and detect structural breaks.
7. **Bayesian Methods:**
   - Utilizes Bayesian techniques to model uncertainty and assess parameter stability by comparing posterior distributions across different time periods or data subsets.

8. **Rolling Regression:**

- Involves estimating the model over consecutive subsets of data, allowing for the observation of changes in coefficients and model fit.

9. **Outlier Analysis:**

- Identifies and analyzes outliers in the data, as these may indicate structural breaks or changes in the underlying data generation process.

10. **Bootstrap Methods:**

- Uses resampling techniques to estimate parameter uncertainty, comparing estimates and confidence intervals across different bootstrap samples.

11. **Context and Domain Knowledge:**

- Structural stability testing should be complemented by a deep understanding of the specific context and domain knowledge to interpret results accurately.

12. **Correction for Multiple Testing:**

- Given the potential for false positives, it's important to correct for multiple testing when conducting various structural stability tests.

13. **Overall Considerations:**

- The choice of a specific test depends on the data characteristics, suspected sources of instability, and the nature of the regression model.

In summary, testing the structural stability of regression models involves a diverse set of statistical techniques that aim to identify changes in relationships over time or across subsets of data, ensuring the reliability and generalizability of the model.

12.7. GLOSSARY

Economic Forecasting: Economic forecasting is the process of making predictions about the future state of an economy, typically involving the analysis and interpretation of various economic indicators, trends, and data. This

forecasting is crucial for policymakers, businesses, investors, and individuals to make informed decisions.

Financial Modeling: Financial modeling is the process of creating a representation of a financial situation or a business using mathematical techniques and tools. Financial models are used for various purposes, including financial analysis, valuation, planning, and decision-making. These models can range from simple spreadsheet calculations to complex computerized simulations

Business Forecasting: Business forecasting is the process of making predictions or estimates about future business conditions, trends, and performance based on historical data, analysis of current factors, and consideration of various influencing variables. Forecasting is a crucial aspect of business planning and decision-making, enabling organizations to anticipate challenges, allocate resources effectively, and capitalize on opportunities

Financial Modeling: Financial modeling is a quantitative analysis technique used to represent the financial performance of a business, project, or investment using mathematical models and tools. Financial models are valuable for decision-making, strategic planning, and assessing the impact of various scenarios on financial outcomes.

12.8 SELF ASSESSMENT QUESTIONS

1. What is Economic Forecasting?

_____

_____

_____

2. Explain the concept of Financial Modeling?

_____

_____

_____

3. Explain the types of testing structural stability of regression models?

_____

_____

_____

4. Uses of testing structural stability of regression models?

_____

_____

_____

12.9 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell*. *Andrew W. Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*

10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"

11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"

12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"

13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

****

UNIT-III **LESSON NO. 13**

REGRESSION WITH QUALITATIVE VARIABLE

DUMMY VARIABLE TRAP

STRUCTURE

13.1 INTRODUCTION

The dummy variable trap is a common issue in statistical modeling, particularly in regression analysis, when using categorical variables to represent different groups or categories. It arises when one or more dummy variables can be predicted perfectly from the others, leading to multicollinearity.

Here's a brief explanation of the dummy variable trap:

1. **Dummy Variables:**
   - In regression analysis, categorical variables with two or more categories are often represented using dummy variables.
   - A dummy variable is binary (0 or 1) and is used to indicate the presence or absence of a particular category.
   - If you have a categorical variable with "k" categories, you typically use "k-1" dummy variables.
2. **Dummy Variable Trap:**
   - The dummy variable trap occurs when there is a perfect linear relationship among the dummy variables.
   - In other words, one dummy variable can be predicted exactly from the others.
   - For example, if you have two dummy variables representing the categories "A" and "B," and you use both in a regression model, the model might become singular or unstable if "A + B = 1" for all observations.
3. **Consequences:**
   - The presence of the dummy variable trap can lead to multicollinearity, a situation where two or more independent variables in a regression model are highly correlated.
   - Multicollinearity can cause problems in estimating the coefficients and their standard errors, making the results less reliable and interpretable.
4. **Avoiding the Dummy Variable Trap:**
   - To avoid the dummy variable trap, always exclude one dummy variable from the model when you have a categorical variable with more than two categories.
   - This exclusion acts as a reference category, and the coefficients of the

remaining dummy variables represent the difference between each category and the reference category.

For example, if you have a variable "Color" with categories "Red," "Green," and "Blue," you can create dummy variables like "IsGreen" and "IsBlue," excluding the dummy for "IsRed."

By addressing the dummy variable trap, you ensure that your regression model is well-defined and doesn't suffer from multicollinearity issues.

13.2 OBJECTIVES

It seems there might be a slight confusion in the phrasing of your question. The "dummy variable trap" is not something with objectives; rather, it's a situation that one aims to avoid in statistical modeling, specifically in regression analysis. Let me clarify:

1. **Avoiding Redundancy:**
   - The primary goal related to the dummy variable trap is to avoid redundancy or perfect multicollinearity in a regression model. When dummy variables are created to represent categorical variables, having a perfect linear relationship (summing to a constant) among them can lead to numerical instability and unreliable estimates.
2. **Ensuring Model Stability:**
   - The presence of a dummy variable trap can make the estimation of coefficients in a regression model unstable. By addressing this issue, one aims to ensure that the model is well-defined and that the estimated coefficients are reliable and have reasonable interpretations.
3. **Interpretability of Coefficients:**
   - The use of dummy variables allows for the inclusion of categorical variables in regression models, providing a way to assess the impact of different categories on the dependent variable. Avoiding the dummy variable trap ensures that the coefficients of these variables are interpretable and reflect meaningful differences between categories.

4. Statistical Validity:

- The dummy variable trap is a concern for statistical validity. When it occurs, the standard errors of the coefficients may become extremely high, making it difficult to draw meaningful conclusions from the statistical analysis. By addressing the trap, one aims to maintain the statistical validity of the regression results.

In summary, the objective is not to fall into the dummy variable trap by structuring the dummy variables in a way that avoids perfect multicollinearity. This allows for a stable and interpretable regression model, ensuring reliable estimates and valid statistical inferences.

13.3 IMPORTANCE OF DUMMY VARIABLES TRAP

It seems there might be some confusion in the phrasing of your question. The "dummy variable trap" itself is not considered important; rather, avoiding the dummy variable trap is crucial in statistical modeling, especially in regression analysis. Let's discuss the importance of addressing the dummy variable trap:

1. Avoiding Multicollinearity:

- The primary importance of addressing the dummy variable trap lies in avoiding multicollinearity, a situation where independent variables in a regression model are highly correlated. Perfect multicollinearity occurs when one dummy variable can be predicted exactly from the others, leading to numerical instability in the estimation process.

2. Ensuring Model Stability:

- The presence of the dummy variable trap can make the estimation of regression coefficients unstable. By avoiding the trap, one ensures that the regression model is well-posed and that the estimation process is numerically stable. This is important for obtaining reliable and consistent estimates.

3. Interpretable Coefficients:

- Using dummy variables allows the inclusion of categorical variables in regression models. Avoiding the dummy variable trap ensures that the coefficients associated with these variables are interpretable. Each coefficient represents the change in the dependent variable relative to the reference category, providing meaningful insights.

4. Statistical Validity:

- Addressing the dummy variable trap is crucial for maintaining the statistical validity of regression results. When multicollinearity is present, the standard errors of coefficients can become inflated, making hypothesis testing unreliable. By avoiding the trap, one ensures that statistical inferences drawn from the regression analysis are valid.

5. Model Robustness:

- A model that falls into the dummy variable trap may not generalize well to new data. By addressing the trap, the model becomes more robust, meaning it is less sensitive to small changes in the data and is more likely to perform well on unseen observations.

In summary, the importance lies in recognizing and avoiding the dummy variable trap to ensure the statistical integrity and stability of regression models, leading to reliable and interpretable results.

13.4 BASICS OF TRAP IN REGRESSION

The term "trap" in regression is often associated with the "dummy variable trap." Let's discuss the basics of the dummy variable trap in the context of regression:

1. Definition:

- The dummy variable trap is a situation in regression analysis where the presence of perfect multicollinearity exists among the independent variables (dummy variables) used to represent categorical variables.

2. **Dummy Variables:**

- Dummy variables are binary (0 or 1) variables used to represent categorical data in regression analysis.
- For a categorical variable with "k" categories, you typically use "k-1" dummy variables. The omitted category serves as the reference or baseline.

3. **Dummy Variable Trap Occurrence:**

- The dummy variable trap occurs when one or more dummy variables can be perfectly predicted from the others. In other words, there is a perfect linear relationship among the dummy variables.

4. **Consequences of Dummy Variable Trap:**

- Multicollinearity: The presence of the dummy variable trap leads to multicollinearity, where two or more independent variables are highly correlated.
- Unreliable Coefficient Estimates: The coefficients associated with the dummy variables become unstable, and their standard errors may become inflated.
- Difficulty in Interpretation: Interpreting the individual effects of dummy variables becomes challenging when multicollinearity is present.

5. **Avoidance Strategies:**

- To avoid the dummy variable trap, one typically excludes one of the dummy variables when modeling categorical variables with more than two categories. This exclusion serves as a reference category, and the coefficients of the remaining dummy variables represent the difference between each category and the reference category.

6. **Example:**

- Suppose you have a categorical variable "Color" with three categories: Red, Green, and Blue. You would create two dummy variables, say "IsGreen" and "IsBlue," and exclude the dummy for "IsRed" to avoid the trap.

7. Importance of Addressing the Trap:

- Addressing the dummy variable trap is crucial for obtaining reliable and interpretable results in regression analysis.
- It ensures that the regression model is well-posed, stable, and the estimated coefficients have meaningful interpretations.

In summary, the dummy variable trap is a common issue in regression modeling, especially when dealing with categorical variables. Recognizing and addressing the trap is essential for ensuring the validity and stability of regression results.

13.5 REGRESSION WITH DUMMY DEPENDENT VARIABLES

The use of dummy variables in regression analysis typically involves having dummy variables as independent variables to represent categorical data. However, if you're referring to a scenario where the dependent variable is binary (taking on values of 0 or 1), it's more common to use logistic regression instead of linear regression.

Let's discuss both scenarios:

1. Linear Regression with Dummy Independent Variables:

- In linear regression, the dependent variable is continuous, and the model assumes a linear relationship between the independent variables and the dependent variable.
- Dummy variables are often used in linear regression to represent categorical data. Each dummy variable corresponds to a category, and its coefficient represents the change in the dependent variable relative to the reference category.
- For example, if you have a categorical variable "Color" with three categories (Red, Green, Blue), you might create two dummy variables (IsGreen and IsBlue) and exclude one (e.g., IsRed) to avoid the dummy variable trap.

2. Logistic Regression with Binary Dependent Variable:

- If your dependent variable is binary (e.g., 0 or 1, Yes or No), logistic regression is more appropriate than linear regression.
- Logistic regression models the probability of the dependent variable taking on the value of 1. The logistic function (S-shaped curve) is used to map a linear combination of the independent variables to probabilities between 0 and 1.
- Dummy variables can still be used as independent variables in logistic regression to represent categorical predictors.

Example of logistic regression with a binary dependent variable:

pythonCopy code

import statsmodels.api as sm # Assuming 'y' is the binary dependent variable and 'X' contains the independent variables (including dummy variables) logit_model = sm.Logit(y, X) result = logit_model.fit() # Print summary statistics print(result.summary())

In summary, if you have a binary dependent variable, logistic regression is a more suitable choice. If you're dealing with a continuous dependent variable and want to incorporate categorical predictors, linear regression with dummy variables is appropriate, with caution to avoid the dummy variable trap.

13.6 IMPORTANCE OF REGRESSION WITH DUMMY DEPENDENT VARIABLES

Regression models with dummy dependent variables are not commonly used or appropriate in standard statistical practice, as they can lead to various issues. In standard regression analysis, the dependent variable is typically continuous, and the models aim to predict or explain its variation based on one or more independent variables. However, there are specialized models that deal with binary or categorical dependent variables, such as logistic regression or multinomial regression.

If by "dummy dependent variables" you mean a binary outcome (0 or

1), let's discuss the importance of using logistic regression or other models designed for such scenarios:

1. **Dealing with Binary Outcomes:**
   - Logistic regression is specifically designed for situations where the dependent variable is binary. It models the probability of an event occurring, making it suitable for scenarios like predicting whether a customer will buy a product (yes/no) or whether a patient has a particular medical condition (positive/negative).
2. **Probability Interpretation:**
   - Logistic regression provides probabilities rather than raw predictions. This is valuable when you are interested in understanding the likelihood of an event happening, especially in fields like medicine, finance, and social sciences.
3. **Handling Non-linearity:**
   - Logistic regression models the relationship between the independent variables and the log-odds of the dependent variable. This allows for capturing non-linear relationships more effectively than a linear regression model with a binary dependent variable.
4. **Avoiding Heteroscedasticity:**
   - Logistic regression helps avoid issues related to heteroscedasticity (unequal variance of errors) that may occur when modeling binary outcomes with linear regression. The assumptions of constant variance and normal distribution of errors are better satisfied in logistic regression.
5. **Interpretability:**
   - The coefficients in logistic regression models can be interpreted in terms of odds ratios, providing insights into the direction and magnitude of the effects of the independent variables on the odds of the event occurring.

6. Robustness:

- Logistic regression tends to be more robust when dealing with binary outcomes and is less sensitive to outliers compared to linear regression with binary dependent variables.

In conclusion, using logistic regression or other models designed for binary outcomes is essential when dealing with dependent variables that are not continuous. These models are tailored to the characteristics of binary data, offering better interpretation and addressing issues that may arise when applying linear regression to non-continuous outcomes.

13.7 USES OF REGRESSION WITH DUMMY DEPENDENT VARIABLES

Regression models with dummy dependent variables, while less common than models with continuous dependent variables, can find application in specific scenarios. It's important to note that such models might have limitations and assumptions that need careful consideration. Here are some potential uses for regression with dummy dependent variables:

1. Binary Outcome Modeling:

- When the dependent variable is binary (e.g., 0 or 1, Yes or No), logistic regression is the standard choice. However, in some cases, researchers might use a linear probability model, a type of regression with a dummy dependent variable, to estimate the effect of predictors on the probability of an event occurring. While this approach is less common due to potential issues, it may be used in certain contexts.

2. Simplification for Interpretation:

- In some cases, researchers might use a dummy dependent variable to simplify interpretation or communication of results. However, it's crucial to acknowledge the limitations and potential pitfalls associated with such an approach.

3. **Ordinal Dependent Variables:**

- In some cases where the dependent variable has ordinal categories (ordered categories with a clear ranking), researchers might use a type of regression with a dummy dependent variable. However, ordinal logistic regression or ordinal probit models are more appropriate for such situations.

4. **Sensitivity Analysis:**

- In sensitivity analyses or robustness checks, researchers might explore different model specifications, including models with dummy dependent variables, to assess the stability and consistency of results.

5. **Exploratory Data Analysis:**

- In the early stages of data exploration, researchers might use different modeling approaches to understand the relationship between variables. This could include exploring models with dummy dependent variables as part of a broader analysis strategy.

It's essential to emphasize that using regression with dummy dependent variables has some drawbacks, including assumptions related to homoscedasticity, normality of errors, and the nature of the dependent variable itself. Logistic regression or other models designed for binary outcomes (such as probit models) are generally preferred when dealing with binary or categorical dependent variables, as they provide more appropriate estimates and handle the non-continuous nature of the dependent variable more effectively. Always carefully consider the characteristics of your data and the assumptions of different modeling approaches when choosing a regression strategy.

13.8 SUMMARY

Regression with dummy dependent variables typically refers to a type of regression analysis where the dependent variable (the variable being predicted) is binary or categorical. In this context, the dependent variable can only take on two possible outcomes, often coded as 0 or 1. This type of regression is commonly known as logistic regression. Here's a summary:

1. **Binary Outcome:** The dependent variable in regression with dummy dependent variables is binary, meaning it can only have two possible values, often representing the presence or absence of an event or the occurrence of a certain category.
2. **Logistic Regression:** Logistic regression is the statistical method used for modeling the relationship between the binary dependent variable and one or more independent variables. Unlike linear regression, which predicts continuous outcomes, logistic regression predicts the probability of the dependent variable belonging to a particular category.
3. **Logit Function:** Logistic regression uses the logistic function (also called the sigmoid function) to model the relationship between the independent variables and the log odds of the dependent variable being 1. The logistic function ensures that the predicted probabilities fall between 0 and 1.
4. **Interpretation of Coefficients:** The coefficients in logistic regression represent the change in the log odds of the dependent variable being 1 associated with a one-unit change in the corresponding independent variable, while holding other variables constant.
5. **Odds Ratio:** The odds ratio is commonly used to interpret the effect of an independent variable in logistic regression. It represents the ratio of the odds of the event occurring in one group to the odds of the event occurring in another group.
6. **Maximum Likelihood Estimation:** Logistic regression estimates the coefficients using maximum likelihood estimation, a statistical method that maximizes the likelihood of observing the given set of outcomes under the assumed model.
7. **Applications:** Logistic regression is widely used in various fields, including medicine, social sciences, finance, and marketing, for tasks such as predicting the likelihood of a disease, the success of a marketing campaign, or the occurrence of an event.

In summary, regression with dummy dependent variables, often referred to as logistic regression, is a statistical approach used when the outcome variable is binary. It models the probability of an event occurring based on one or more predictor variables and is particularly useful in situations where linear regression is not appropriate due to the nature of the dependent variable.

13.7 GLOSSARY

Dummy Variables: Dummy variables, also known as indicator variables or binary variables, are used in statistical modeling and regression analysis to represent categorical data with two or more categories in a quantitative manner. These variables are called "dummy" because they take on the values 0 or 1, serving as a way to encode categorical information into a format that can be used in statistical models.

Regression with dummy dependent variables: It seems there might be a misunderstanding in your question. Regression models with dummy dependent variables are relatively uncommon because regression models are typically used when the dependent variable is continuous. Dummy variables (indicator variables) are more commonly used for categorical independent variables or factors.

If you are dealing with a binary outcome variable (i.e., a variable that can take only two values, often coded as 0 or 1), logistic regression is the more appropriate model to use. Logistic regression models the probability that the dependent variable belongs to a particular category, given the values of the independent variables.

Dummy Variable Trap: The dummy variable trap is a situation that occurs when two or more dummy variables in a regression model are highly correlated. This correlation can cause issues in the estimation of coefficients and may lead to multicollinearity, a condition where independent variables are highly correlated with each other. The dummy variable trap specifically refers to the scenario where one dummy variable can be predicted perfectly from the others in the model. This perfect multicollinearity can interfere with the

estimation process and make it impossible to obtain unique estimates for the coefficients.

13.8 SELF ASSESSMENT QUESTIONS

1. What is meaning of basics of trap in regression?

_____

_____

_____

2. Explain the Importance of regression with dummy dependent variables?

_____

_____

_____

3. Explain the regression with dummy dependent variables?

_____

_____

_____

13.11 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati

6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University

7. Hill R.C., Griffiths W.E., Lim G.C. “Principles of Econometrics”, 4th edition, [ISBN 978-1-11803207-*7*

8. Jeffrey M. Wooldridge “Introductory Econometrics A Modern Approach”, *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107

9. *John Y. Campbell*. *Andrew W*. *Lo*. A. *Craig MacKinlay,* 1996 “The Econometrics of Financial Markets” ISBN: 9780691043012; Published: *Dec 29*, *1996*

10. Robert H. Shumway, David S. Stoffer” Characteristics of Time Series”

11. Robert H. Shumway, David S. Stoffer “Time Series Regression and Exploratory Data Analysis”

12. Robert H. Shumway, David S. Stoffer “Statistical Methods in the Frequency Domain”

13. Robert H. Shumway , David S. Stoffer “Time Series Analysis and Its ApplicationsWith R Examples”

14. Aileen Nielsen, 2019“Practical Time Series Analysis” Publisher(s): O’Reilly Media, Inc. ISBN: 9781492041658.

15. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ

16. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, “Introduction to Time Series Analysis and Forecasting”

17. Greene, William H., Econometric Analysis, Prentice Hall, 2000.

18. Econometrics by Damodar Gujarati

19. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University

20. Hill R.C., Griffiths W.E., Lim G.C. “Principles of Econometrics”, 4th edition, [ISBN 978-1-11803207-*7*

21. Jeffrey M. Wooldridge “Introductory Econometrics A Modern Approach”, *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107

22. John Y. Campbell. Andrew W. Lo. *A*. Craig MacKinlay, *1996* “The Econometrics of Financial Markets” ISBN: 9780691043012; Published: *Dec 29*, *1996*

23. Robert H. Shumway, David S. Stoffer” Characteristics of Time Series”

24. Robert H. Shumway, David S. Stoffer “Time Series Regression and Exploratory Data Analysis”

25. Robert H. Shumway, David S. Stoffer “Statistical Methods in the Frequency Domain”

26. Robert H. Shumway , David S. Stoffer “Time Series Analysis and Its ApplicationsWith R Examples”

27. Ruud, Paul A., 2000 “An Introduction to Classical Econometric Theory”, Oxford, 2000.

28. Wooldridge, J M, 2009 “Introductory Econometrics – A Modern Approach” (4th ed), South-Western, 2009.

****

UNIT-III **LESSON NO. 14**

REGRESSION WITH QUALITATIVE VARIABLE

DUMMY VARIABLE TRAP

STRUCTURE

14.1 INTRODUCTION

In econometrics, "LMP" commonly stands for "Limited Dependent Variable Model," and "logit" typically refers to the logistic regression model. The combination "LMP model logit" often refers to a specific type of econometric model used when the dependent variable is binary or categorical with limited outcomes.

Here's a breakdown of the components:

1. **Limited Dependent Variable Model (LMP):** This type of model is employed when the dependent variable is restricted in some way, such as being binary (taking on only two possible values, like 0 or 1) or categorical with a limited number of outcomes. Examples of limited dependent variables include whether a person is employed (1) or unemployed (0) or whether a customer makes a purchase (1) or not (0).

2. **Logit Model:** The logit model is a type of regression model used for binary or dichotomous dependent variables. It transforms the probability of an event occurring into a log-odds scale, making it more amenable to linear regression techniques. The logistic regression model is mathematically represented by the logistic function (sigmoid function), and it models the log-odds of the dependent variable being in a particular category.

   Combining these concepts, "LMP model logit" suggests a limited dependent variable model that uses logistic regression. This is often applied when dealing with situations where the dependent variable is binary or categorical, and the logistic function is used to model the relationship between the independent variables and the probability of the outcome.

   It's worth noting that the logit model is just one approach, and there are other methods for modeling limited dependent variables, such as probit regression. The choice between logit and probit models often depends on the specific characteristics of the data and the assumptions made about the underlying distribution of the error terms.

14.2. OBJECTIVES OF LMP MODEL LOGIT

The Limited Dependent Variable Model (LMP) with a logit specification in econometrics serves several objectives when analyzing data with binary or categorical dependent variables. Here are some of the primary objectives:

1. Modeling Binary or Categorical Outcomes:

- **Binary Outcomes:** The logit model is particularly useful when the dependent variable is binary, meaning it can take only two possible outcomes (e.g., success/failure, employed/unemployed).
- **Categorical Outcomes:** It can also be extended to handle categorical outcomes with more than two categories through techniques like multinomial logistic regression.

2. Estimating Probabilities:

- The logit model estimates the probabilities associated with each category of the dependent variable. It models the log-odds (logit) of an event occurring, and by applying the logistic function, these log-odds are transformed into probabilities.

3. Handling Non-Linearity:

- The logistic function introduces a non-linear element to the model, allowing it to capture complex relationships between independent and dependent variables.

4. Dealing with Heteroscedasticity:

- The logit model helps address issues related to heteroscedasticity, which is the situation where the variability of the error term is not constant across observations. The logistic transformation can stabilize variance.

5. Interpretability:

- The coefficients in a logit model represent the change in the log-odds of the dependent variable for a one-unit change in the independent variable. These log-odds can be converted into odds ratios, providing a more

interpretable measure of the impact of independent variables on the probability of an event.

6. Handling Endogeneity:

- The logit model can help mitigate issues related to endogeneity by providing consistent estimates when certain assumptions are met. Endogeneity arises when independent variables are correlated with the error term.

7. Model Comparison:

- The logit model allows for comparison with alternative models, such as probit models, to determine which specification better fits the data based on likelihood ratio tests, Wald tests, or other statistical criteria.

8. Prediction and Classification:

- The logit model can be used for prediction and classification tasks, such as predicting the probability of an event or classifying observations into different categories based on their estimated probabilities.

In summary, the objectives of employing an LMP model with a logit specification in econometrics include providing a flexible framework for modeling limited dependent variables, estimating probabilities, addressing non-linear relationships, and offering interpretability for the relationships between independent and dependent variables.

14.3 GROUPED LOGIT MODEL

The Grouped Logit Model is an extension of the standard Logit Model in econometrics, specifically designed to handle situations where the data are grouped or clustered. This model is particularly useful when observations are not independent but are grouped in some way, and there is potential correlation or heterogeneity within groups. Let's explore the key features and applications of the Grouped Logit Model:

1. Grouped Data:

- The Grouped Logit Model is suitable for situations where data can be grouped into clusters, panels, or some other form of grouping.

- Examples of grouped data could include observations at different time periods, data from different geographic regions, or data from different experimental units.

2. Correlated Observations:

- In scenarios where observations within a group are likely to be correlated or share common unobserved characteristics, the standard Logit Model may not account for this correlation.
- The Grouped Logit Model allows for correlation within groups, providing a more accurate representation of the underlying structure of the data.

3. Hierarchical Structure:

- The model is suitable for hierarchical or nested data structures where observations are naturally grouped together.
- For example, in a study involving students within schools, or patients within hospitals, the data can be structured hierarchically, and the Grouped Logit Model can capture the correlation within each group.

4. Random Effects:

- The Grouped Logit Model often incorporates random effects, which account for unobserved heterogeneity within groups.
- Random effects can capture variations in the intercepts or slopes across different groups, allowing for a more flexible and realistic modeling approach.

5. Estimation Techniques:

- Estimation of the Grouped Logit Model is typically done using specialized techniques, such as generalized estimating equations (GEE) or maximum likelihood estimation (MLE).
- These techniques take into account the clustered nature of the data and provide consistent and efficient parameter estimates.

6. Applications:

- The Grouped Logit Model is commonly used in various fields such as health economics, education research, and social sciences where observations are naturally grouped.
- It can be applied to model individual choices within groups or clusters, taking into consideration the potential correlation or heterogeneity within those groups.

In summary, the Grouped Logit Model is a valuable tool in econometrics when dealing with data that exhibit a grouped structure. It provides a more accurate representation of the underlying correlations within groups and allows researchers to account for unobserved heterogeneity at the group level.

14.4 OBJECTIVES OF GROUPED LOGIT MODEL

The Grouped Logit Model is employed in econometrics to achieve specific objectives when dealing with data that exhibit a grouped or clustered structure. Here are some of the primary objectives of using the Grouped Logit Model:

1. Capture Correlation within Groups:

- One of the main objectives of the Grouped Logit Model is to account for the potential correlation or dependence among observations within the same group or cluster.
- By considering this correlation, the model provides more accurate parameter estimates, taking into account the shared characteristics or unobserved factors within each group.

2. Handle Hierarchical Data Structures:

- The model is designed to handle hierarchical or nested data structures where observations are naturally grouped. Examples include students within schools, patients within hospitals, or firms within industries.
- By incorporating random effects or other mechanisms, the Grouped Logit Model accommodates the inherent structure of the data, allowing for a

more realistic representation of the relationships within and between groups.

3. **Account for Unobserved Heterogeneity:**

- Unobserved heterogeneity refers to variations in individual characteristics that are not directly observable. The Grouped Logit Model allows for the inclusion of random effects, which capture this unobserved heterogeneity within groups.
- By considering random effects, the model accounts for differences in the intercepts or slopes across different groups, providing a more flexible and nuanced representation of the underlying dynamics.

4. **Improve Efficiency of Parameter Estimates:**

- The Grouped Logit Model, through its specialized estimation techniques such as generalized estimating equations (GEE) or maximum likelihood estimation (MLE), aims to provide consistent and efficient parameter estimates.
- The inclusion of group-specific effects allows the model to better utilize the information contained within each group, leading to more precise parameter estimates.

5. **Enhance Predictive Accuracy:**

- By addressing correlation within groups and incorporating random effects, the Grouped Logit Model often leads to improved predictive accuracy compared to models that do not account for the grouped nature of the data.
- The model's ability to capture both systematic and unobserved variations within and between groups contributes to a more reliable representation of the decision-making process.

6. **Facilitate Group-Specific Inference:**

- Researchers often have specific interest in understanding and making inferences about group-specific effects. The Grouped Logit Model allows

for such group-specific inference by estimating parameters at both the individual and group levels.

In summary, the primary objectives of the Grouped Logit Model revolve around accurately modeling the correlation within groups, accommodating hierarchical data structures, capturing unobserved heterogeneity, improving parameter estimation efficiency, enhancing predictive accuracy, and enabling group-specific inferences. These objectives make the model particularly useful in situations where observations are naturally grouped or clustered.

14.5 IMPORTANCE OF GROUPED LOGIT MODEL

The Grouped Logit Model is a statistical technique used in econometrics and market research to analyze discrete choice data. It is an extension of the multinomial logit model, specifically designed to handle situations where the choices are grouped or clustered in some way. Here are some reasons why the Grouped Logit Model is important:

1. Handling Grouped Data:

- In real-world scenarios, choices are often made in groups or clusters. For example, individuals within a household may make joint decisions on certain purchases. The Grouped Logit Model is well-suited for analyzing such grouped data, allowing for a more accurate representation of decision-making processes.

2. Correlated Choices:

- In situations where choices made by one individual within a group are correlated with the choices of others, the Grouped Logit Model is advantageous. It allows for the modeling of these correlations, providing a more realistic and comprehensive understanding of decision dynamics.

3. Efficiency and Computational Advantages:

- The Grouped Logit Model can offer computational advantages when compared to modeling each choice individually. It allows for more efficient estimation of parameters, especially when dealing with a large number of choices or groups.

4. Reduced Parameter Dimensionality:

- When choices are grouped, the Grouped Logit Model can help reduce the dimensionality of the parameter space. This is beneficial in terms of model estimation and interpretation, making it more feasible to handle complex decision scenarios.

5. Improved Predictive Accuracy:

- By considering the interdependencies within groups, the Grouped Logit Model can provide more accurate predictions of choices. This is particularly important in market research and policy analysis where understanding consumer behavior is crucial for making informed decisions.

6. Policy Implications:

- The insights derived from the Grouped Logit Model can have significant policy implications. For instance, in transportation planning, understanding how households make joint decisions about travel modes can inform the development of more effective transportation policies.

7. Accounting for Heterogeneity:

- The model allows for the incorporation of individual-level and group-level heterogeneity, capturing variations in preferences and decision-making processes. This helps in better reflecting the diversity of choices within and between groups.

8. Market Segmentation:

- The Grouped Logit Model is useful in market segmentation studies, where understanding how different demographic groups or clusters of consumers make choices can aid businesses in tailoring their products and marketing strategies.

In summary, the Grouped Logit Model is important because it addresses the limitations of standard multinomial logit models when dealing with grouped or clustered choices. It provides a more realistic and nuanced approach to

analyzing decision-making processes in various fields, offering improved predictive accuracy and better informing policy and business decisions.

14.6 DESCRIBE THE GROUPED LOGIT MODEL

The Grouped Logit Model is a statistical model used in econometrics and market research to analyze discrete choice data, especially when choices are made in groups or clusters. It is an extension of the multinomial logit model, which is a standard tool for modeling discrete choices.

Here's a description of the key components and concepts of the Grouped Logit Model:

1. Discrete Choices:

- The model deals with situations where individuals or entities must make a choice among two or more mutually exclusive alternatives. These choices are discrete, meaning that individuals can only choose one option from the available set.

2. Multinomial Logit Foundation:

- The Grouped Logit Model builds upon the multinomial logit framework. In a multinomial logit model, the probability of choosing a particular alternative is modeled as a function of explanatory variables, and the coefficients of these variables are estimated through maximum likelihood estimation.

3. Grouped Choices:

- In the Grouped Logit Model, the choices are organized into groups or clusters. This is particularly relevant when choices are made jointly by individuals within a group, such as households making decisions on purchases or activities.

4. Correlated Choices:

- The model allows for correlations among choices within a group. This is crucial in situations where the decisions made by one individual may

be correlated with the decisions of others within the same group. Accounting for these correlations improves the model's ability to represent real-world decision-making processes.

5. **Group-Specific Parameters:**

- The Grouped Logit Model introduces group-specific parameters to account for variations in preferences or decision processes between groups. Each group may have its own set of parameters, capturing the heterogeneity in choices across different clusters.

6. **Likelihood Function:**

- Similar to the multinomial logit model, the estimation of parameters in the Grouped Logit Model is typically done through maximizing the likelihood function. The likelihood function represents the probability of observing the actual choices given the model parameters.

7. **Estimation Techniques:**

- Maximum likelihood estimation is commonly used to estimate the parameters of the Grouped Logit Model. This involves finding the values of the parameters that maximize the likelihood of observing the given choices.

8. **Policy Implications and Predictions:**

- Once the model is estimated, it can be used to make predictions about the likelihood of different choices within and between groups. These predictions have implications for policy decisions, market segmentation, and understanding the impact of various factors on decision outcomes.

In summary, the Grouped Logit Model extends the multinomial logit framework to better capture the complexities of decision-making when choices are grouped or correlated. It is a valuable tool in various fields, including transportation planning, market research, and policy analysis, where understanding group dynamics is essential for making informed decisions.

14.7 APPLICATION OF THE GROUPED LOGIT MODEL

The Grouped Logit Model, an extension of the standard multinomial logit model, is particularly useful in situations where choices are made in groups or clusters. Here are some common applications of the Grouped Logit Model:

1. Transportation Planning:

- Analyzing travel behavior within households or groups, such as joint decisions on travel modes or destinations. The model can account for correlations in choices made by individuals within the same group.

2. Market Research:

- Studying consumer choices within households or families, especially for products or services that are likely to be chosen jointly. For example, understanding the preferences of family members when selecting vacation destinations or purchasing a car.

3. Health Economics:

- Investigating healthcare decisions made by families or social groups, such as the choice of health insurance plans or treatment options. The model can capture correlations in health-related choices within a household.

4. Education Economics:

- Analyzing educational choices within households, such as decisions related to school enrollment or participation in educational programs. The Grouped Logit Model can be applied to understand the joint decision-making process within families.

5. Housing and Real Estate:

- Modeling housing choices within households, such as decisions about housing types, locations, or rental versus ownership. The model can capture correlations in housing preferences within families.

6. **Policy Analysis:**

- Assessing the impact of policy interventions on group-level choices. For example, understanding how certain policies influence joint decisions within communities or social groups.

7. **Environmental Economics:**

- Analyzing group-level choices related to environmental conservation or sustainable practices. The model can be used to understand how environmental decisions are made within households or communities.

8. **Social Sciences:**

- Investigating various social decisions and behaviors within groups, such as recreational choices, cultural activities, or participation in community events. The model can capture correlations in these decisions.

9. **Tourism and Hospitality:**

- Studying travel and accommodation choices made by groups, such as families or friends traveling together. The Grouped Logit Model can provide insights into joint decision-making in the tourism sector.

10. **Family Economics:**

- Understanding economic decisions within families, such as joint choices related to expenditures, savings, or investment decisions. The model can account for correlations in financial decisions within households.

11. **Energy Economics:**

- Analyzing group-level decisions related to energy consumption and conservation practices within households or communities. The model can capture correlations in energy-related choices.

12. **Public Health Interventions:**

- Assessing the impact of public health campaigns or interventions on group-level behaviors, such as smoking cessation or dietary choices within families.

13. Travel Demand Modeling:

- Modeling transportation choices for groups traveling together, such as families or coworkers commuting to work. The model can provide insights into joint decision-making in travel demand studies.

14. Joint Purchases:

- Understanding group-level decisions in joint purchases, such as households deciding on large expenditures like appliances or furniture.

The Grouped Logit Model is valuable in situations where choices are interdependent within groups, and understanding these interdependencies is crucial for accurate modeling and analysis. It provides a more realistic representation of decision-making processes in various fields.

14.8 CONCEPT OF PROBIT MODEL IN ECONOMETRICS

The Probit Model is a statistical model used in econometrics to analyze binary outcome variables, which take on only two possible values, often denoted as 0 and 1. The term "Probit" is a contraction of "probability unit," emphasizing the focus on modeling probabilities. The Probit Model is particularly useful when the dependent variable represents a binary choice or event, and it provides a way to link explanatory variables to the probability of observing a particular outcome. Here are the key concepts associated with the Probit Model:

1. Binary Outcome:

- The Probit Model is suitable for situations where the dependent variable is binary, meaning it can only take on two possible values. Common examples include whether an individual buys a product (1 for yes, 0 for no), whether a student passes an exam (1 for pass, 0 for fail), etc.

2. Probability Function:

- The model assumes that the probability of observing the outcome of interest is related to a linear combination of explanatory variables through a standard normal cumulative distribution function (CDF), usually denoted Φ (phi).

3. Link Function :

- The link function in the Probit Model is the inverse of the standard normal CDF. Mathematically, it is represented as: $(=1) = \Phi(0 + 11 + 22 + ...+)P(Yi = 1) + \Phi(\beta 0 + \beta 1X1i + \beta 2X2i + ...\beta kXki)$ where $(=1) = P(Yi = 1)$ is the probability of the binary outcome being 1 for the i-th observation, Φ Φ is the standard normal CDF, $0\beta 0$ is the intercept, $1, 2...., \beta 1, \beta 2....\beta k$ are the coefficients associated with the explanatory variables 1,2,…,X1i, X2i,…,Xki.

4. Interpretation of Coefficients:

- The coefficients in the Probit Model represent the change in the z-score (standardized normal deviate) for a one-unit change in the corresponding explanatory variable. The z-score is then transformed into a probability using the standard normal CDF.

5. Maximum Likelihood Estimation (MLE):

- The Probit Model is typically estimated using Maximum Likelihood Estimation. The objective is to find the values of the coefficients that maximize the likelihood of observing the given set of outcomes based on the assumed model.

6. Predictions:

- Once the model is estimated, it can be used to predict the probability of the binary outcome for new observations or to assess the impact of changes in the explanatory variables on the probability of the outcome occurring.

7. Goodness of Fit:

- The goodness of fit of the Probit Model can be assessed through various statistical tests and measures, such as the likelihood ratio test or the Hosmer-Lemeshow test.

The Probit Model is widely used in econometrics, especially when dealing with binary outcomes in fields like labor economics, health economics,

and finance. It provides a flexible framework for modeling the relationship between explanatory variables and the probability of an event occurring, taking into account the normal distribution of errors.Top of Form

14.9 OBJECTIVES OF PROBIT MODEL IN ECONOMETRICS

The Probit Model in econometrics serves several objectives, providing a valuable tool for analyzing binary outcome variables. Here are the key objectives and applications of the Probit Model:

1. Modeling Binary Outcomes:

- The primary objective of the Probit Model is to model and analyze binary outcome variables, where the response variable can take on only two possible values (usually coded as 0 and 1). Common examples include choices, decisions, or events such as whether an individual purchases a product, defaults on a loan, or passes an exam.

2. Probability Estimation:

- The Probit Model is designed to estimate the probability of the occurrence of the binary outcome based on a set of explanatory variables. It provides a way to quantify the likelihood of an event happening given certain conditions.

3. Understanding the Impact of Explanatory Variables:

- The coefficients in the Probit Model represent the impact of explanatory variables on the probability of the binary outcome. By examining these coefficients, researchers can understand the direction and magnitude of the influence of each variable.

4. Policy Analysis:

- The Probit Model is often used in policy analysis to assess the impact of policy interventions or changes in economic conditions on the likelihood of a specific event or decision. This can inform policymakers about the potential effectiveness of different policy options.

5. Risk Assessment and Management:

- In various fields, such as finance and insurance, the Probit Model is used for risk assessment and management. It helps in predicting the likelihood of certain events, such as loan defaults or insurance claims, and can aid in decision-making related to risk mitigation.

6. Labor Economics:

- In labor economics, the Probit Model is frequently employed to analyze labor market outcomes, such as the probability of being employed, the likelihood of union membership, or the probability of job turnover.

7. Health Economics:

- In health economics, the Probit Model is used to analyze binary health outcomes, such as the likelihood of a patient responding positively to a treatment, the probability of adopting healthy behaviors, or the presence of a medical condition.

8. Evaluating Program Impact:

- Researchers use the Probit Model to assess the impact of interventions or programs on binary outcomes. This is common in fields like education, where researchers may evaluate the effectiveness of educational programs on student outcomes.

9. Market Research:

- In market research, the Probit Model can be applied to analyze consumer choices and predict the likelihood of adopting a new product or brand based on various marketing factors.

10. Statistical Inference:

- The Probit Model allows for hypothesis testing and statistical inference to determine the significance of individual coefficients and overall model fit. This helps in assessing the reliability of the estimated relationships.

In summary, the Probit Model is a versatile tool in econometrics, providing a framework for modeling and analyzing binary outcomes. Its

applications extend across various fields, offering insights into decision-making processes, policy effects, and the impact of explanatory variables on the probability of specific events.

14.10 APPLICATION OF PROBIT MODEL IN ECONOMETRICS

The Probit Model finds applications in various areas of econometrics and beyond, where the analysis of binary outcomes is crucial. Here are some common applications of the Probit Model:

1. Labor Economics:

- Analyzing factors influencing employment decisions, such as the probability of being employed or the likelihood of participating in the labor force.

2. Health Economics:

- Studying health-related outcomes, such as the probability of adopting healthy behaviors, the likelihood of seeking medical treatment, or the probability of a patient responding positively to a treatment.

3. Finance:

- Assessing the likelihood of loan default, bankruptcy, or financial distress. The Probit Model is also used in credit scoring to predict the probability of an individual defaulting on a loan.

4. Marketing Research:

- Evaluating consumer choices and preferences, such as the probability of purchasing a product, subscribing to a service, or switching brands.

5. Education Economics:

- Analyzing educational outcomes, like the probability of a student passing an exam, completing a degree, or participating in specific educational programs.

6. **Criminal Justice:**

- Predicting the likelihood of criminal behavior, recidivism, or compliance with parole conditions.

7. **Public Policy Analysis:**

- Assessing the impact of policy interventions on binary outcomes, such as the probability of individuals participating in government programs or complying with regulations.

8. **Market Segmentation:**

- Understanding consumer segmentation by predicting the probability of individuals belonging to specific market segments based on their preferences and behaviors.

9. **Insurance:**

- Predicting the probability of insurance claims or the likelihood of specific events covered by insurance policies.

10. **Environmental Economics:**

- Analyzing binary environmental outcomes, such as the likelihood of households adopting eco-friendly practices or complying with environmental regulations.

11. **Human Resources:**

- Assessing employee-related outcomes, like the probability of employee turnover, job satisfaction, or engagement in workplace programs.

12. **Political Science:**

- Studying voting behavior, such as the likelihood of individuals voting for a particular candidate or party.

13. **Econometric Analysis of Experiments:**

- Evaluating the impact of experimental treatments or interventions on binary outcomes in controlled studies.

14. Biostatistics and Epidemiology:

- Analyzing health-related outcomes, such as the probability of disease occurrence, response to a treatment, or the presence of specific health conditions.

15. Quality Control:

- Predicting the likelihood of defects or failures in manufacturing processes.

The Probit Model, along with its counterpart, the Logit Model, provides a flexible framework for modeling and analyzing binary outcomes. Its applications span a wide range of disciplines, making it a versatile tool in empirical research and decision-making.

14.11 CONCEPT OF TOBIT MODEL IN ECONOMETRICS

The Tobit Model is a statistical model used in econometrics to analyze data where the dependent variable is observed with censoring or truncation. Censoring occurs when some observations in the data set have values that fall below or above a certain threshold and are recorded as equal to that threshold. The Tobit Model was introduced by James Tobin in 1958.

Here are the key concepts associated with the Tobit Model:

1. Censoring:

- The Tobit Model is specifically designed to handle censored data. Censoring occurs when the dependent variable is not fully observed due to limitations in measurement instruments or because values fall outside a certain range. Censoring is often referred to as left-censoring (values below a certain threshold), right-censoring (values above a certain threshold), or interval-censoring.

2. Dependent Variable:

- The dependent variable in the Tobit Model is latent, meaning it is not directly observed. Instead, what is observed is a censored version of the

latent variable. The latent variable is assumed to have a linear relationship with explanatory variables.

3. Linear Structural Equation:

- The Tobit Model posits a linear structural equation for the latent variable: ?=+y β=X β+u where β y β is the latent variable, X is the matrix of explanatory variables, β is the vector of coefficients, and u is the error term.

4. Censoring Mechanism:

- The observed dependent variable (y) is related to the latent variable through a censoring mechanism. The censored values are determined by the threshold and are observed as: =max?(min?(?,),)yi? =max(min(yi β,U),L) where yi is the observed dependent variable for observation i, U is the upper censoring threshold, and L is the lower censoring threshold.

5. Probability Density Function:

- The Tobit Model assumes that the latent variable follows a normal distribution. The probability density function (PDF) for the latent variable is expressed as: $()+1(-)\int(u)=\sigma(1\phi(\sigma u-X\beta)$ where ϕ is the standard normal density function, σ is the standard deviation of the error term, and u is the error term.

6. Maximum Likelihood Estimation (MLE):

- The parameters of the Tobit Model, including the coefficients and the variance of the error term, are estimated using Maximum Likelihood Estimation (MLE). The likelihood function is constructed based on the observed censored data.

7. Interpretation of Coefficients:

- The coefficients in the Tobit Model represent the marginal effects of the explanatory variables on the expected value of the latent variable. These can be interpreted similarly to coefficients in linear regression models.

8. Heteroscedasticity:

- The Tobit Model assumes homoscedasticity, meaning constant variance of the error term across all observations. Heteroscedasticity can be a concern, and adjustments may be needed.

The Tobit Model is commonly used in economics and social sciences when dealing with censored data, such as income data that may have a lower censoring threshold (e.g., zero income) or upper censoring threshold (e.g., a certain income level). It provides a way to model and estimate relationships when the dependent variable is not fully observed due to censoring.

14.12 OBJECTIVES OF TOBIT MODEL IN ECONOMETRICS

The Tobit Model in econometrics serves several objectives, addressing specific challenges associated with censored or truncated data. Here are the key objectives of using the Tobit Model:

1. Handling Censored Data:

- The primary objective of the Tobit Model is to handle situations where the dependent variable is censored, meaning some values are not observed because they fall below or above a certain threshold. This is particularly common in economic data, such as income data with a lower censoring threshold (e.g., zero income) or an upper censoring threshold.

2. Accounting for Limited Observability:

- The Tobit Model allows for the analysis of data where the dependent variable is only partially observable due to censoring. By modeling the latent variable and incorporating the censoring mechanism, it provides a way to estimate relationships even when some values are not directly observed.

3. Estimating Latent Variables:

- The model aims to estimate the latent variable, which represents the unobservable, continuous variable of interest. This latent variable is assumed to have a linear relationship with explanatory variables.

4. **Incorporating Censoring Mechanism:**

- The Tobit Model explicitly incorporates the censoring mechanism into the modeling framework. It recognizes that observed values are censored versions of the true, unobserved values and provides a probabilistic representation of this process.

5. **Maximum Likelihood Estimation (MLE):**

- The Tobit Model uses Maximum Likelihood Estimation (MLE) to estimate the parameters of the model. MLE aims to find the values of the parameters that maximize the likelihood of observing the censored data given the model.

6. **Interpreting Coefficients:**

- The coefficients in the Tobit Model represent the marginal effects of the explanatory variables on the expected value of the latent variable. These coefficients can be interpreted similarly to coefficients in linear regression models.

7. **Understanding Determinants of Censoring Thresholds:**

- In cases where there are both upper and lower censoring thresholds, the Tobit Model can provide insights into the determinants of these thresholds. For example, understanding factors that influence the minimum or maximum values that can be observed in the data.

8. **Handling Corner Solutions:**

- The Tobit Model is suitable for handling "corner solutions" where the dependent variable is constrained at one or both ends (e.g., zero income). This is a common scenario in economic data, and the Tobit Model provides a way to account for such constraints.

9. **Assessing Model Fit:**

- The Tobit Model allows for the assessment of how well the model fits the observed censored data. Various diagnostic tests can be applied to evaluate the goodness of fit and the appropriateness of the model assumptions.

In summary, the Tobit Model is a valuable tool in econometrics for addressing the challenges associated with censored or truncated data. Its objectives include estimating latent variables, modeling the censoring mechanism, and providing insights into the determinants of censoring thresholds, making it particularly useful in the analysis of economic and social data with limited observability.

14.13 IMPORTANCE OF TOBIT MODEL IN ECONOMETRICS

Importance of the tobit model in econometrics ChatGPT

ChatGPT

The Tobit Model is an important tool in econometrics, offering a valuable framework for analyzing data with censoring or truncation. Its significance lies in its ability to handle situations where the dependent variable is only partially observed, providing insights into relationships that may be obscured by censoring. Here are some key reasons why the Tobit Model is important in econometrics:

1. Handling Censored Data:

- The primary importance of the Tobit Model is its ability to handle censored data. In various economic and social contexts, data limitations or constraints may lead to censoring of observations, and the Tobit Model provides a systematic way to account for this censoring.

2. Modeling Corner Solutions:

- The Tobit Model is suitable for situations where the dependent variable is constrained at one or both ends (corner solutions). For example, when dealing with income data, the model can address the issue of left-censoring at zero.

3. Estimation of Latent Variables:

- The Tobit Model allows for the estimation of latent variables, representing the unobservable, continuous variable of interest. This is

particularly relevant when the variable of interest is only partially observable due to censoring.

4. **Insights into Determinants of Censoring Thresholds:**
   - In cases where there are both upper and lower censoring thresholds, the Tobit Model can provide insights into the determinants of these thresholds. Understanding what factors influence the minimum or maximum values that can be observed is valuable in economic analysis.
5. **Efficient Use of Censored Data:**
   - By explicitly modeling the censoring mechanism, the Tobit Model makes efficient use of censored data. It incorporates information from censored observations, improving the precision of parameter estimates and enhancing the overall efficiency of the estimation process.
6. **Economic Interpretability:**
   - The Tobit Model allows for the economic interpretation of coefficients, similar to linear regression models. This facilitates the understanding of how changes in explanatory variables affect the latent variable and, consequently, the observed censored variable.
7. **Application in Labor Economics:**
   - In labor economics, where earnings data often exhibit censoring at zero, the Tobit Model is frequently applied to analyze factors influencing income levels and understand the determinants of labor market outcomes.
8. **Application in Health Economics:**
   - In health economics, the Tobit Model can be used to study healthcare expenditure data, where values are often left-censored at zero. It helps in understanding the factors influencing healthcare spending.
9. **Improved Inference and Prediction:**
   - The Tobit Model provides a more accurate representation of the underlying relationship in censored data, leading to improved inference and prediction compared to models that do not account for censoring.

10. Accounting for Heteroscedasticity:

- The Tobit Model assumes homoscedasticity (constant variance of the error term). However, extensions of the model, such as the heteroscedastic Tobit model, can be applied to handle situations where the variance of the error term may vary across observations.

In summary, the Tobit Model is important in econometrics due to its ability to handle censored data, provide insights into latent variables, and offer a systematic approach to modeling situations where the dependent variable is only partially observable. Its applications extend across various economic and social domains, making it a valuable tool in empirical research.

14.14 APPLICATION OF TOBIT MODEL IN ECONOMETRICS

The Tobit Model finds applications in various areas of econometrics where the dependent variable is censored or limited, leading to challenges in data analysis. Here are some common applications of the Tobit Model in econometrics:

1. Income Analysis:

- Analyzing income data, where many observations may be censored at zero, especially for individuals with no income. The Tobit Model helps in understanding the factors influencing income levels and how different variables contribute to income variation.

2. Labor Economics:

- Investigating factors affecting labor market outcomes, such as earnings, wages, or hours worked. The Tobit Model is particularly useful when dealing with earnings data that may have a substantial number of observations censored at zero.

3. Expenditure Analysis:

- Studying expenditure patterns in households or individuals, where certain expenditures may be censored at zero. The model can provide insights

into factors influencing spending behavior and the determinants of zero expenditures.

4. **Health Economics:**
   - Analyzing healthcare expenditure data, where individuals may have zero expenditures due to not utilizing healthcare services. The Tobit Model helps in understanding the factors affecting healthcare spending patterns.

5. **Research and Development Expenditure:**
   - Investigating research and development expenditures in firms, where firms that do not engage in R&D activities may have expenditures censored at zero. The Tobit Model can shed light on the determinants of R&D spending.

6. **Environmental Economics:**
   - Studying environmental conservation efforts and expenditures, where some individuals or firms may have zero expenditures on environmentally friendly practices. The Tobit Model can be applied to understand the factors influencing environmental spending.

7. **Marketing Research:**
   - Analyzing consumer spending behavior and purchase decisions, especially for products with a significant number of non-purchasers. The Tobit Model can help identify factors affecting the likelihood and amount of spending.

8. **Education Economics:**
   - Investigating education-related expenditures or investments, such as spending on books, tutoring, or educational resources. The Tobit Model can be applied to understand the determinants of educational spending.

9. **Innovation and Patent Analysis:**
   - Analyzing innovation and patent-related data, where some firms may have zero patent counts. The Tobit Model can be used to examine the factors influencing innovation and patenting behavior.

10. Savings and Investment Analysis:

- Studying savings and investment patterns, especially when dealing with financial data where some individuals may have zero savings or investment amounts. The Tobit Model helps in understanding factors influencing financial decisions.

11. Analysis of Program Participation:

- Investigating factors influencing participation in government assistance programs or social welfare programs. The Tobit Model can be applied to analyze the impact of various factors on program participation.

12. Survey Data Analysis:

- Analyzing survey data where certain responses are censored due to constraints or limitations. The Tobit Model provides a framework for modeling and understanding the determinants of censored survey responses.

These applications highlight the versatility of the Tobit Model in handling censored data across various economic and social domains. The model's ability to provide insights into latent variables and analyze relationships in the presence of censoring makes it a valuable tool in econometric research.

14.15 SUMMARY

The Tobit model is a statistical model commonly used in econometrics for handling censored dependent variables, where the observations are not fully observed but are instead censored or truncated. Here's a summary of the Tobit model in econometrics:

1. **Censored Data:** The Tobit model is designed to address situations where the dependent variable is censored, meaning that some observations fall below or above a certain threshold and are not directly observable. This often occurs when the dependent variable is restricted to a specific range.
2. **Two-Part Model:** The Tobit model is a two-part model. The first part models the probability of observing a value above or below the censoring

threshold, often using a probit or logit model. The second part models the level of the dependent variable for those observations that are not censored, typically using a linear regression model.

3. **Threshold Parameter:** The Tobit model includes a threshold parameter that represents the cutoff point below or above which the dependent variable is censored. This parameter is estimated from the data.

4. **Assumptions:** The Tobit model assumes that the errors in both parts of the model are normally distributed. This is crucial for obtaining efficient estimates through maximum likelihood estimation.

5. **Interpretation:** The coefficients in the Tobit model are interpreted in a similar way to those in linear regression for the uncensored part of the data. They represent the marginal effect of a one-unit change in the independent variable on the expected value of the dependent variable.

6. **Applications:** The Tobit model is commonly used in economics and social sciences when dealing with data that has censoring or truncation issues. Examples include modeling household expenditures, income, or working hours, where there might be lower or upper bounds.

7. **Challenges:** Tobit models assume a normal distribution for the errors, and violations of this assumption can impact the validity of the results. Additionally, the model may be sensitive to the choice of the censoring threshold.

In summary, the Tobit model is a valuable tool in econometrics for analyzing censored data. It combines a probabilistic approach for modeling the likelihood of observing censored values with a regression approach for estimating the relationship between the independent and dependent variables for uncensored observations.

14.16 GLOSSARY

Logit Model: The Logit Model is a type of regression model used for predicting the probability of a binary outcome. It is a specific form of logistic

regression, which is designed for situations where the dependent variable is categorical and has only two possible values (0 and 1). The logistic function, also known as the sigmoid function, is used in the logit model to transform a linear combination of predictor variables into a probability.

The logistic function (denoted by *P*) is defined as follows:

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

Here:

- P(Y=1) is the probability that the dependent variable Y takes the value 1.
- $\beta 0, \beta 1 .... \beta n$ are coefficients to be estimated.
- *X1, X2,...,Xn* are the independent variables.

The logit model is derived from the logistic function by taking the natural logarithm (log-odds transformation) of the odds ratio:

$$\left(1 - P(Y=1)P(Y=1)\right) = \beta 0, \beta 1 X 1, \beta 2 X 2 + ... + \beta n Xn$$

In this form, the left side of the equation represents the log-odds of the probability of the dependent variable being 1.

Grouped Logit Model: The Grouped Logit Model is an extension of the standard Logit Model, designed to handle situations where the data can be grouped into clusters or shares. In this model, individuals within the same group or share are assumed to have similar tastes or preferences, leading to a shared response function.

The Grouped Logit Model is particularly useful when dealing with choice or decision data where individuals are grouped into clusters, and the outcome of interest is binary. It is commonly used in areas such as transportation, marketing, and social sciences to analyze choices made by individuals within distinct groups.

The basic form of the Grouped Logit Model can be expressed as follows:

$$\ln\left(\frac{P(Y_{ij}=1)}{1-P(Y_{ij}=1)}\right) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_n X_{ijn}$$

where:

- Yij is the binary outcome for individual i in group j.
- P(Yij=1) is the probability of the outcome being *1* for individual *i* in group *j*.
- $\beta 0, \beta 1 .... \beta n$ are coefficients to be estimated.
- *Xij1, Xij2, ...,Xijn* are the independent variables for individual *i* in group *j*.

Tobit Model: The Tobit Model is a statistical model that is used when the dependent variable is censored or truncated. Censoring occurs when the values of the dependent variable are observed up to a certain threshold, and values beyond that threshold are not observed. The Tobit Model accounts for both the observed values and the censored values in the analysis.

The Tobit Model is named after James Tobin, who introduced it in 1958. It is commonly used in econometrics and social sciences, especially in situations where the dependent variable has a substantial proportion of observations at a specific limit or threshold, and there is interest in modeling both the observed and censored portions of the data.

The Tobit Model can be specified as follows:

$$yi+ = \beta 0 + \beta 1 xi1 + \beta 2 xi2 + ... \beta kxik + ui$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

where:

- *yi* is the observed dependent variable (censored at 0).

- yi^* is the latent variable, representing the unobservable true value of the dependent variable.
- xi1,xi2...xik are the independent variables.
- $\beta 0, \beta 1 .... \beta n$ are coefficients to be estimated.
- *ui* is the error term.

14.17 SELF ASSESSMENT QUESTIONS

1. What is the meaning of Grouped logit model?

_____

_____

_____

2. Explain the Importance of LMP model logit?

_____

_____

_____

3. How is the Probit model useful in econometrics?

_____

_____

_____

14.18 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"

4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.

5. Econometrics by Damodar Gujarati

6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University

7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*

8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107

9. *John Y. Campbell*. *Andrew W. Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*

10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"

11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"

12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"

13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.

15. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ

16. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"

17. Greene, William H., Econometric Analysis, Prentice Hall, 2000.

18. Econometrics by Damodar Gujarati

19. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University

20. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*

21. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107

22. John Y. Campbell. Andrew W. Lo. *A*. Craig MacKinlay, *1996* "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*

23. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"

24. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"

25. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"

26. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

27. Ruud, Paul A., 2000 "An Introduction to Classical Econometric Theory", Oxford, 2000.

28. Wooldridge, J M, 2009 "Introductory Econometrics – A Modern Approach" (4th ed), South-Western, 2009.

****

UNIT-III **LESSON NO. 15**

REGRESSION WITH QUALITATIVE VARIABLE

MODELING COUNT DATA AND POISSON MODEL

STRUCTURE

15.1 INTRODUCTION ABOUT MODELING COUNT DATA

Modeling count data refers to the statistical analysis and development of models to understand and predict outcomes that are represented by non-negative integer counts. Count data arise in various fields where the focus is on the number of occurrences of events within a specific time period, space, or category. The modeling process involves selecting an appropriate statistical model that accounts for the discrete and non-negative nature of the data.

Here are key points regarding the meaning of modeling count data:

1. **Type of Data:** Count data consist of whole numbers that represent the frequency of events. Examples include the number of accidents at an intersection, the count of customer arrivals in a store, the number of phone calls received in a day, or the occurrences of a particular disease in a population.

2. **Statistical Models:** Various statistical models are used for modeling count data. One of the common models is Poisson regression, which assumes that events occur independently at a constant rate. Other models, such as negative binomial regression, zero-inflated models, or hurdle models, may be employed depending on the characteristics of the data.

3. **Poisson Regression:** Poisson regression is a widely used model for count data. It assumes that the count variable follows a Poisson distribution, and it models the logarithm of the mean count as a linear combination of predictor variables.

4. **Overdispersion:** Overdispersion occurs when the variance of the count data is greater than the mean, violating the assumption of equality in the Poisson distribution. In such cases, alternative models like negative binomial regression are often considered.

5. **Applications:** Count data modeling is applicable in a variety of fields, including epidemiology, finance, biology, social sciences, and engineering. For example, in epidemiology, one might model the count of disease cases based on various risk factors.

6. **Interpretation of Results:** The results of count data models are interpreted in terms of the impact of predictor variables on the count outcome. Coefficients in the model represent the change in the expected count associated with a one-unit change in the corresponding predictor, while accounting for other variables.

7. **Model Evaluation:** Model evaluation involves assessing the goodness of fit to ensure that the chosen model adequately represents the underlying patterns in the count data. Diagnostic tools, such as residual analysis and goodness-of-fit tests, are commonly used.

In summary, modeling count data is a statistical process that involves selecting an appropriate model to understand the relationship between predictor variables and the count outcome. The goal is to develop a model that accurately captures the patterns and variability in the count data for effective prediction and interpretation.

15.2 OBJECTIVES OF MODELING COUNT DATA

Modeling count data involves developing statistical models to understand and predict outcomes that are represented by non-negative integers, such as the number of events or occurrences within a specific time period, space, or category. The objectives of modeling count data can be summarized as follows:

1. **Understanding Relationships:** Identify and understand the relationships between predictor variables and the count outcome. Determine how changes in the values of predictors are associated with changes in the count of events.

2. **Prediction:** Develop a model that can be used for predicting future counts based on the values of predictor variables. This is particularly useful for making informed decisions and planning in various fields.

3. **Risk Assessment:** Assess the risk factors that contribute to the occurrence of events represented by count data. Identify variables that are statistically significant in influencing the count outcome.

3. **Causal Inference:** Investigate potential causal relationships between predictor variables and the count outcome. Understand the impact of different factors on the frequency of events.

4. **Comparisons and Contrasts:** Compare and contrast the effects of different predictor variables on the count outcome. Assess which variables have a more substantial impact and whether certain conditions lead to significant changes in the count.

5. **Model Interpretability:** Develop a model that is interpretable and provides insights into the relationships between predictors and the count outcome. Interpretability is crucial for communicating findings to a broader audience and making informed decisions.

6. **Model Selection:** Choose an appropriate statistical model that best fits the characteristics of the count data. Common models include Poisson regression, negative binomial regression, zero-inflated models, or hurdle models, depending on the nature of the data.

7. **Accounting for Overdispersion:** Address overdispersion, which occurs when the variance of the count data is greater than the mean. Select models or techniques that account for overdispersion, such as negative binomial regression, to obtain more accurate and reliable results.

8. **Model Validation:**

8.1. Validate the chosen model to ensure that it provides a good fit to the observed count data. Use diagnostic tools, such as residual analysis and goodness-of-fit tests, to assess the model's performance.

9. **Decision Support:** Provide support for decision-making processes by offering insights into the factors influencing count outcomes. The model results can guide actions or interventions to mitigate risks or enhance positive outcomes.

10. **Scientific Inquiry:** Contribute to scientific inquiry and knowledge by gaining a deeper understanding of the underlying processes generating

the count data. This is particularly relevant in research fields where count data play a significant role.

In summary, the objectives of modeling count data encompass gaining insights, making predictions, understanding relationships, and providing valuable information for decision-making and scientific inquiry in diverse fields.

15.3 ADVANTAGES OF MODELING COUNT DATA

Modeling count data using appropriate statistical methods offers several advantages in various fields. Here are some key advantages of modeling count data:

1. **Suitability for Discrete Data:** Count models are specifically designed for situations where the outcome variable is discrete and represents the number of occurrences of events. Traditional regression models, which assume continuous outcomes, may not be well-suited for count data.
2. **Interpretability:** Count models, such as Poisson regression, provide interpretable results. The coefficients can be exponentiated to obtain multiplicative effects, making it easy to understand the impact of predictor variables on the rate of occurrence of events.
3. **Accounting for Variability:** Count models account for the inherent variability in count data. The Poisson distribution, for example, assumes that events occur independently at a constant rate, and the variance is equal to the mean.
4. **Handling Overdispersion:** Overdispersion, where the variance is greater than the mean, is common in count data. Models like negative binomial regression can handle overdispersed data, providing a more flexible alternative to Poisson regression.
5. **Applicability in Diverse Fields:** Count data models find applications in a wide range of fields, including epidemiology, finance, ecology, criminology, social sciences, and more. They can be tailored to the specific characteristics of the data and research questions in these diverse domains.

6. **Risk Assessment and Prediction:** Count models are valuable for assessing risks associated with specific events. By identifying relevant predictors, these models can be used to predict the likelihood of future events, helping in risk management and decision-making.

7. **Model Flexibility:** Various count models, such as zero-inflated models or hurdle models, offer flexibility in handling different patterns within count data, such as excess zeros or two-part processes where certain conditions must be met for an event to occur.

8. **Scientific Inquiry:** Count data models contribute to scientific inquiry by helping researchers understand the relationships between variables and the underlying processes generating the count data. This can lead to the development of new theories or the refinement of existing ones.

9. **Decision Support:** The insights gained from count data models can support decision-making processes in areas such as public health, finance, and marketing. Understanding the factors influencing event occurrences can inform strategies and interventions.

10. **Availability of Statistical Software:** Statistical software packages (e.g., R, Python with libraries like **statsmodels** or **scikit-learn**) provide tools for fitting and evaluating count models, making them accessible to researchers and analysts.

11. **Robustness to Outliers:** Count models are often robust to the presence of outliers in the response variable, making them suitable for data with extreme values.

In summary, modeling count data offers advantages in terms of addressing the unique characteristics of discrete outcomes, providing interpretable results, and supporting decision-making across various disciplines. The choice of the specific count model depends on the nature of the data and the research questions at hand.

15.4 TYPES OF MODELING COUNT DATA

There are several types of models designed specifically for modeling

count data. The choice of a particular model depends on the characteristics of the data and the research questions of interest. Here are some common types of models for modeling count data:

1. **Poisson Regression:** The Poisson regression model is one of the most commonly used models for count data. It assumes that the count variable follows a Poisson distribution, and it models the logarithm of the mean count as a linear combination of predictor variables.
2. **Negative Binomial Regression:** Negative binomial regression is an extension of the Poisson regression model that accounts for overdispersion, where the variance of the count data is greater than the mean. This model introduces an additional parameter to handle the extra variability.
3. **Zero-Inflated Models:** Zero-inflated models are used when there is an excess of zeros in the count data compared to what would be expected from a Poisson or negative binomial distribution. These models combine a binary process determining whether an observation is zero or not, along with a count distribution for the non-zero counts.
4. **Hurdle Models:** Hurdle models are similar to zero-inflated models but assume that the zero counts are generated by a separate process (a hurdle), and then a count distribution is used for the non-zero counts. These models are suitable when there is a structural barrier that prevents events from occurring.
5. **Zero-Truncated Models:** Zero-truncated models exclude observations with zero counts from the analysis. These models are used when zero counts are impossible or not relevant. For example, if counting starts only after the occurrence of the first event.
6. **Poisson Process Models:** Poisson process models are used when events are assumed to occur randomly in time or space. These models are suitable for temporal or spatial count data.
7. **Geometric Distribution Models:** The geometric distribution represents

the number of trials needed for a success in a sequence of Bernoulli trials. Geometric models can be applied to count data when modeling the number of trials until the first success.

8. **Poisson-Tweedie Models:** Poisson-Tweedie models are a generalization of Poisson regression that includes the Poisson distribution as a special case and allows for a range of variance structures, making it more flexible for handling different types of count data.

9. **Compound Poisson Models:** Compound Poisson models consider a two-stage process where the occurrence of an event follows a Poisson distribution, and the magnitude of the event follows a different distribution. These models are useful for situations where the size of events varies.

10. **Generalized Linear Mixed Models (GLMMs):** GLMMs extend traditional count models by incorporating random effects, making them suitable for handling clustered or correlated count data.

The choice of a specific model depends on the nature of the count data, the presence of excess zeros, overdispersion, and other characteristics unique to the dataset. It's important to assess model fit and compare different models to choose the one that best represents the underlying process generating the count data.

15.5 CONCEPT ABOUT THE POISSON REGRESSION:

Poisson Regression is a statistical model used for analyzing count data. It is particularly well-suited for situations where the outcome variable represents the number of events or occurrences within a fixed unit of time or space. The model is named after the French mathematician Siméon Denis Poisson, who introduced the Poisson distribution.

Key Characteristics and Components of Poisson Regression:

1. **Nature of the Outcome Variable:**

- Poisson regression is designed for count data, which consists of non-

negative integers representing the number of events or occurrences. Examples include the number of accidents at an intersection, the count of customer arrivals in a store, or the number of phone calls received in a day.

2. Probability Distribution:

- The Poisson distribution is the underlying probability distribution for the count variable. It assumes that events occur independently and at a constant average rate within a fixed interval. The probability mass function of the Poisson distribution is given by P(X=k)=k!e? $\lambda\lambda$ k, where λ is the average rate of events.

3. Link Function:

- The Poisson regression model employs a log-link function, transforming the mean of the Poisson distribution into a linear combination of predictor variables. The relationship is expressed as log $(\lambda) = \beta 0 + \beta 1x1 + 1\beta 2 + 2 + ... + \beta kxk$, where λ is the expected count, log is the natural logarithm, and $\beta 0, \beta 1, ...\beta k$ are the coefficients.

4. Assumption of Equality of Mean and Variance:

- Poisson regression assumes that the mean and variance of the count variable are equal. This assumption may be violated in the presence of overdispersion, where the variance is greater than the mean. In such cases, negative binomial regression or other models may be considered.

5. Maximum Likelihood Estimation (MLE):

- The parameters (coefficients) of the Poisson regression model are estimated using maximum likelihood estimation (MLE). The goal is to find the parameter values that maximize the likelihood of observing the given count data under the assumed Poisson distribution.

6. Interpretation of Coefficients:

- The coefficients (β) in Poisson regression are exponentiated to obtain multiplicative effects on the rate of the count variable. For example, if

the coefficient for a predictor is 0.1, it implies a 10% increase in the expected count for a one-unit increase in that predictor, holding other variables constant.

7. Goodness of Fit:

- Model assessment involves evaluating the goodness of fit, typically through residual analysis, goodness-of-fit tests, and diagnostic plots. Deviance is often used as a measure of goodness of fit.

8. Overdispersion

- If overdispersion is detected, where the variance exceeds the mean, negative binomial regression or other overdispersed count models may be considered as alternatives.

Poisson regression is widely used in fields such as epidemiology, biology, finance, and social sciences to model and analyze count data. It provides a simple and interpretable framework for understanding the relationship between predictor variables and the frequency of events.

15.6 POISSON MODEL ANALYSIS

Poisson model

- The Poisson model predicts the number of occurrences of an event.
- The Poisson model states that the probability that the dependent variable Y will be equal to a certain number y is:

$$p(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- For the Poisson model, ↵ is the intensity or rate parameter.

$$\mu = \exp\left(\mathbf{x}_i^{'}\beta\right)$$

- Interpretation of the coefficients: one unit increase in x will increase/decrease the average

number of the dependent variable by the coefficient expressed as a percentage.

Properties of the Poisson distribution

- Equidispersion property of the Poisson distribution: the equality of mean and variance.

$$E(y|x) = var(y|x) = \mu$$

This is a restrictive property and often fails to hold in practice, i.e., there is "overdispersion" in the data. In this case, use the negative binomial model.

- Excess zeros problem of the Poisson distribution: there are usually more zeros in the data than a Poisson model predicts. In this case, use the zero-inflated Poisson model.

Marginal effects for the Poisson model

- The marginal effect of a variable on the average number of events is:

$$\partial E(y|x)/\partial x_j = \beta_j \exp(\mathbf{x}_i'\beta)$$

- Interpretation of the marginal effects: one unit increase in x will increase/ decrease the average number of the dependent variable by the marginal effect.

Negative binomial model

- The negative binomial model is used with count data instead of the Poisson model if there is Over dispersion in the data.

- Unlike the Poisson model, the negative binomial model has a less restrictive property that the variance is not equal to the mean

$$var(y|x) = \mu + \alpha\mu^2$$

- Another functional form is $var(y|x) = \mu + \alpha\mu,$, but this form is less used.

- The negative binomial model also estimates the overdispersion parameter Ù.

Test for overdispersion

- Estimate the negative binomial model which includes the over dispersion parameter *t* and test
- if *t* is significantly different than zero.
- H0: a=0 or Ha: a=0
- We have three cases:
- When a = 0, the Poisson model.
- When a > 0, over dispersion (frequently holds with real data).
- When a < 0, under dispersion (not very common).

Incidence rate ratios (irr)

- For the Poisson and negative binomial models, in addition to reporting the coefficients and marginal effects, we can also report the incidence rate ratios.
- The incidence rate ratios report exp(b) rather than b.
- Interpretation of the incidence rate ratios: irr=2 means that for each unit increase in x, the expected number of y will double.

Hurdle or two-part models

- The two-part model relaxes the assumption that the zeros (whether or not there are events) and positives (how many events) come from the same data generating processes.
- Example: different factors may affect whether or not you practice a particular sport and how many times you practice your sport in a month.

- We can estimate two-part models similar to the truncated regression models.
- If the process generating the zeros is $f_1(.)$ and the process generating the positive responses is $f_2(.)$ then the two-part hurdle model is defined by the following probabilities:

$$g(y) = \begin{cases} f_1(0) \ \ if \ y = 0 \\ \dfrac{1 - f_1(0)}{1 - f_2(0)} f_2(y) \ \ if \ y \geq 1 \end{cases}$$

- If the two processes are the same, this is the standard count data model.
- The model for the zero versus positive responses is a binary model with the specified distribution, but we usually estimate it with the probit/logit model.

Zero-inflated models

- The zero-inflated model is used with count data when there is an excess zeros problem.
- The zero-inflated model lets the zeros occur in two different ways: as a realization of the binary process (z=0) and as a realization of the count process when the binary variable z=1.
- Example: you either like hiking or you do not. If you like hiking, the number of hiking trips you can take is 0, 1, 2, 3, etc. So you may like hiking, but may not take a trip this year. We are able to generate more zeros in the data.
- If the process generating the zeros is f1(.) and the process generating the positive responses is f2(.) then the zero-inflated model is:

$$g(y) = \begin{cases} f_1(0) + (1 - f_1(0)) f_2(0) \ \ if \ y = 0 \\ (1 - f_1(0)) f_2(y) \ \ if \ y \geq 1 \end{cases}$$

- The zero-inflated model is less frequently used than the hurdle model.
- The zero-inflated models can handle the excess zeros problem.

15.7 POISSON REGRESSION:

The Poisson regression is a statistical model that is used to analyze count data, where the outcome variable represents the number of events or occurrences within a fixed unit of time or space. Named after the French mathematician Siméon Denis Poisson, this regression model is particularly suited for situations where the data are discrete and non-negative.

Key Concepts of Poisson Regression:

1. **Count Data:** The Poisson regression is designed for count data, which consists of non-negative integers representing the number of times an event occurs in a specific time period or spatial region.

2. **Rate or Intensity:** The underlying assumption is that events occur randomly and independently at a constant rate within the specified interval. The Poisson distribution describes the probability of observing a particular number of events in a fixed interval.

3. **Link Function:** In Poisson regression, the relationship between the mean of the count variable variable (λ , the average rate of occurrence) and the predictor variables is modeled using a log-linear relationship.The link function is the natural logarithm (log), and the model is often expressed as $\log(\lambda) = \beta 0 + \beta 1 x 1 + 1\beta 2 + 2 + ... + \beta ixj$.

4. **Exponential Family Distribution:** Poisson regression is a member of the exponential family of distributions, which provides a broad framework for modeling various types of outcomes. It falls under the category of generalized linear models (GLMs).

5. **Maximum Likelihood Estimation (MLE):** The parameters (β coefficients) of the Poisson regression model are estimated using maximum likelihood estimation (MLE). The goal is to find the parameter values that maximize the likelihood of observing the given set of count data, given the predictor variables.

6. **Interpretation of Coefficients:** The coefficients in a Poisson regression

model are exponentiated to obtain the multiplicative effect on the rate of the outcome variable. For example, if the coefficient for a predictor is 0.1, it implies that a one-unit increase in that predictor is associated with an approximately 10% increase in the expected count, holding other variables constant.

7. **Goodness of Fit:** Model assessment in Poisson regression involves evaluating the goodness of fit. Common methods include examining residuals, conducting goodness-of-fit tests, and checking for overdispersion.

8. **Overdispersion:** Poisson regression assumes that the variance of the count variable is equal to its mean. If the data exhibit greater variability than expected under this assumption (overdispersion), alternative models such as negative binomial regression may be considered.

In summary, the Poisson regression model provides a statistical framework for analyzing count data by modeling the rate of events in relation to predictor variables. It is widely used in fields such as epidemiology, finance, biology, and social sciences to understand and quantify the impact of covariates on the frequency of events.

15.8 ADVANTAGES OF POISSON REGRESSION

What are the advantages of Poisson regression?

Poisson regression offers several advantages when it comes to modeling count data. Here are some of the key advantages:

1. **Applicability to Count Data:** Poisson regression is specifically designed for modeling count data, where the outcome variable represents the number of occurrences of an event in a fixed interval of time or space. It is a natural choice for situations where the data are discrete and non-negative.

2. **Interpretability:** The coefficients in a Poisson regression model have a straightforward interpretation. For example, if the estimated coefficient

for a predictor variable is 0.1, it implies a multiplicative effect: a one-unit increase in the predictor is associated with an approximately 10% increase in the expected count, holding other variables constant.

3. **Simple Model Structure:** The simplicity of the Poisson regression model makes it easy to implement and understand. The model assumes a constant rate of event occurrence, and the estimation process is relatively straightforward.

4. **Well-Developed Statistical Theory:** Poisson regression is based on well-established statistical theory, and there are efficient methods for estimating model parameters. Maximum likelihood estimation (MLE) is commonly used, and statistical tests and confidence intervals are readily available.

5. **Suitability for Rare Events:** Poisson regression is particularly well-suited for situations where the events being counted are rare. If the mean rate of occurrence is low, Poisson distribution often approximates the distribution of the count variable.

6. **Availability in Statistical Software:** Poisson regression is widely supported in statistical software packages, making it easily accessible for researchers and analysts. Popular statistical programming languages like R and Python provide libraries for fitting Poisson regression models.

However, it's important to note that Poisson regression also has limitations. One key assumption is that the mean and variance of the count variable are equal, which may not hold in all cases. In situations where overdispersion is present (variance > mean), negative binomial regression or other models may be more appropriate. Additionally, if there are excess zeros in the data, alternative models like zero-inflated or hurdle models might be considered.

15.9 EXAMPLES OF POISSON REGRESSION

Poisson regression is commonly used in various fields to model count data. Here are a few examples where Poisson regression might be applied:

1. **Traffic Accidents:** Suppose you want to model the number of traffic accidents at a particular intersection each day. Poisson regression could be used to analyze the relationship between predictor variables (e.g., traffic volume, weather conditions) and the count of accidents.
2. **Healthcare:** In medical research, Poisson regression might be employed to model the number of hospital admissions for a specific condition over time, considering factors such as age, gender, and other relevant covariates.
3. **Epidemiology:** Poisson regression is often used in epidemiology to study the incidence rates of diseases. For example, researchers might use Poisson regression to analyze the number of cases of a particular infectious disease in a population based on various risk factors.
4. **Web Traffic:** In web analytics, Poisson regression could be applied to model the number of page views or clicks on a website over a period, taking into account factors like marketing campaigns, day of the week, or time of day.
5. **Insurance Claims:** Insurance companies might use Poisson regression to model the number of insurance claims filed within a specific time period. Predictors could include policy features, customer demographics, or other relevant variables.
6. **Customer Complaints:** In customer service analysis, Poisson regression could be used to model the number of complaints received by a company each day, considering factors such as product type, service quality, or promotional activities.

Ecology: Ecologists may use Poisson regression to model the number of animal sightings or plant occurrences in a specific habitat, taking into account environmental variables.

Criminal Incidents: Law enforcement agencies might use Poisson regression to analyze the number of criminal incidents in a particular area, considering factors such as population density, socioeconomic status, and policing efforts.

These examples illustrate the versatility of Poisson regression in modeling count data across different domains. Keep in mind that in cases of overdispersion or excess zeros, alternative models like negative binomial regression or zero-inflated models might be more appropriate.

15.10. DIFFERENCE BETWEEN POISSON REGRESSION AND NORMAL REGRESSION

What is the difference between Poisson regression and normal regression?

Poisson regression and normal (linear) regression are two different types of regression models that are used for distinct types of data. Here are the key differences between Poisson regression and normal regression:

1. **Nature of the Outcome Variable:** Poisson Regression: It is used when the outcome variable is a count, representing the number of occurrences of an event in a fixed interval of time or space. The Poisson distribution is often used to model count data, assuming events occur independently and at a constant rate.

   Normal (Linear) Regression: It is used when the outcome variable is continuous and follows a normal distribution. Normal regression assumes that the residuals (the differences between observed and predicted values) are normally distributed.

2. **Assumptions about the Residuals:** Poisson Regression: Assumes that the variance of the outcome variable is equal to its mean, which is a characteristic of the Poisson distribution. It is also designed for count data, which often exhibit a right-skewed distribution.

   Normal (Linear) Regression: Assumes that the residuals are normally distributed and have constant variance across all levels of the predictor variables. This is known as homoscedasticity.

3. **Link Function:** Poisson Regression: Uses a log-link function. The relationship between the predictors and the mean of the Poisson distribution is modeled as a logarithmic function.

Normal (Linear) Regression: Assumes a linear relationship between the predictors and the mean of the normal distribution.

4. **Model Output and Interpretation:** Poisson Regression: The coefficients are exponentiated and represent multiplicative effects on the rate of the outcome variable. Interpretation is often in terms of percentage change in the rate of occurrence.

   Normal (Linear) Regression: The coefficients represent additive effects on the mean of the outcome variable. Interpretation is in terms of units change in the mean.

5. **Handling Outliers:** Poisson Regression: Robust to outliers in the response variable, as it is designed for count data. However, it assumes that the mean and variance are equal, so it may not perform well with overdispersed data.

   Normal (Linear) Regression: Sensitive to outliers, and extreme values in the response variable can disproportionately influence the model parameters.

In summary, the choice between Poisson regression and normal regression depends on the nature of the outcome variable. Poisson regression is appropriate for count data, while normal regression is suitable for continuous data with a normal distribution. Additionally, for count data, it's important to consider alternative models like negative binomial regression if overdispersion is present.

15.11 SUMMARY

Poisson Regression is a statistical model commonly used for count data, where the outcome variable represents the number of events occurring in a fixed interval of time or space. Here's a summary of Poisson Regression:

1. **Count Data:** Poisson Regression is specifically designed for situations where the dependent variable is a count of events or occurrences. It is widely used in fields such as epidemiology, insurance, biology, and criminology, where events happen independently and at a constant rate.

2. **Assumption:** The key assumption of the Poisson model is that the mean and variance of the count variable are equal. This is suitable when events are random and occur independently.

3. **Probability Distribution:** The Poisson distribution is the underlying probability distribution for the Poisson Regression model. It describes the probability of a given number of events occurring in a fixed interval of time or space.

4. **Model Structure:** The Poisson Regression model assumes that the expected count of events (the mean) is a function of one or more predictor variables. Mathematically, it is expressed as $(\lambda) = e\beta 0 + \beta 1x1 + 1\beta 2 + 2 + ...,$ where λ is the expected count, and $...\beta 0, \beta 1,...$ are the coefficients associated with the predictor variables.

5. **Link Function:** The natural logarithm (log) is commonly used as the link function in Poisson Regression, linking the mean count to the linear combination of predictor variables.

6. **Interpretation of Coefficients:** The coefficients in Poisson Regression represent the multiplicative effect of a one-unit change in the predictor variable on the expected count of events, assuming all other variables are held constant.

7. **Overdispersion:** In cases where the assumption of equal mean and variance is violated (overdispersion), Negative Binomial Regression may be used as an alternative to Poisson Regression.

8. **Applications:** Poisson Regression is applied to various scenarios such as modeling the number of accidents, the frequency of disease occurrences, or counts of customer arrivals in a given time period.

In summary, Poisson Regression is a valuable tool for modeling count data, providing insights into the relationships between predictor variables and the expected frequency of events. Its simplicity and interpretability make it particularly useful for situations where count outcomes are prevalent.

15.12 GLOSSARY

Poisson Regression:

Poisson Regression is a type of regression analysis used when the dependent variable is a count or a rate, representing the number of times an event occurs within a fixed period or in a specific area. It is named after the French mathematician Siméon Denis Poisson, who made significant contributions to probability theory.

The Poisson Regression model is appropriate for situations where the count data is non-negative integers (0, 1, 2, ...) and where the mean and variance of the count are assumed to be equal. The model is often used in fields such as epidemiology, biology, criminology, and other disciplines dealing with count data.

The Poisson Regression model can be expressed as follows:

$$\ln(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

where:

- λi is the expected value (mean) of the count variable for observation i.
- ln ?ln is the natural logarithm.
- $\beta 0, \beta 1, ... \beta k$ are coefficients to be estimated.
- xi1,xi2,…,xik are the independent variables.

The Poisson Regression model assumes that the observed count variable *yi* follows a Poisson distribution with mean λi, and the relationship between λi and the independent variables is modeled using a linear combination of predictors.

$$P(Y = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

Count Data: Count data refers to data that represent the number of occurrences or events within a specified unit of observation, typically discrete

and non-negative. This type of data is characterized by whole numbers (integers) and often arises in various fields, including biology, epidemiology, finance, social sciences, and many others. Count data is distinct from continuous data, which can take any value within a range.

Normal regression: It seems there might be a misunderstanding or a broad term used. Generally, when we refer to "normal regression," we are likely talking about linear regression using the normal distribution. Linear regression models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

The basic linear regression model can be expressed as:

$$(Y) = \beta 0 + \beta 1x1 + 1\beta 2 + 2 + ...,+ \beta nXn + \varepsilon$$

Here:

- Y is the dependent variable.
- X1,X2,…,Xn are the independent variables.
- $\beta 0, \beta 1, / dusots.\beta n$ are the coefficients to be estimated.
- ε represents the error term, which captures unobserved factors affecting Y.

15.13 SELF ASSESSMENT QUESTIONS

1. Explain the Types of modeling count data?

_____

_____

_____

2. What is the meaning of Poisson regression?

_____

_____

_____

3. **What is the Difference between poisson regression and normal regression?**

_____

_____

_____

15.14 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell. Andrew W. Lo.* A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29, 1996*
10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"
11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"

12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"

13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.

15. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ

16. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"

17. Greene, William H., Econometric Analysis, Prentice Hall, 2000.

18. Econometrics by Damodar Gujarati

19. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University

20. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*

21. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107

22. *John Y. Campbell. Andrew W. Lo. A. Craig MacKinlay, 1996* "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*

23. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"

24. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"

25. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"

26. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

27. Ruud, Paul A., 2000 "An Introduction to Classical Econometric Theory", Oxford, 2000.

28. Wooldridge, J M, 2009 "Introductory Econometrics – A Modern Approach" (4th ed), South-Western, 2009.

****

UNIT-IV **LESSON NO. 16**

TIME SERIES ECONOMETRICS

TIME SERIES ANALYSIS - BASIC, UTILITY AND SERIES OF TIME

STRUCTURE

16.1 INTRODUCTION

In today's modern and digital world, statisticians are pretty much occupied with analyzing consumer patterns. We are generating a huge amount

of data, which should be simply trashed. There is a tremendous amount of value to the data generated. If processed correctly, it can gain fortunes for the organization by preparing it to the mindset of its consumers. Whether we want to assess the consumers' electricity consumption pattern or study the statistics behind the financial trends in the market, time analysis plays a crucial role.

In the modern world, where there is a huge importance on technological research and booming digital technology, time is an essential factor that needs to be considered. To predict consumer usage analysis which can be financial investments, electricity consumption, expenditure on e-commerce, or predicting the positive growing stocks in the future and the planning, the investment, etc., time series plays a crucial role.

In definition terms, a time series is generally a series of ordered points on the timeline, with time always being the independent variable to predict the future trend.

The data can be from one of the three types:

- **Time Series Data:** This is nothing but the noted or observational values taken at different time frames.
- **Cross-Sectional Data:** Data from one or more dependent variables collected at the same given time.
- **Pooled Data:** This is hybrid data which can be a combination of data and cross-sectional data.

Mathematically the time series can be obtained by the below equation:

y=f(t), with t being the independent variable time, and y is generally the response to the time over the function.

16.2 OBJECTIVES

Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur. With modern analytics platforms, these visualizations can go far beyond line graphs.

16.3 WHY DO WE NEED TIME SERIES ?

- A series of events indexed based on time is Time Series.
- They are mostly plotted using line graphs or line charts.
- To answer why we need time series, we need to know the vast area where they are implemented. This list will be extensive as prediction is becoming a major influencing factor for organizations to garnish their consumers.
- It has its fundamentals in statistics and probability, so statisticians widely employ it.
- It is also helpful for the digital signal processor, where we often see time as one of the independent variables.
- Pattern recognition basing one of some predefined characteristics is one of the applications where it has identified its presence. Also, time series is vastly helpful for mathematicians in econometrics.
- It has founded its application in earthquake detection, estimating impacted areas during the prediction of natural calamities, and understanding weather patterns over time.
- Apart from the abovementioned fields, it also has found its application in astronomy, control engineering, and electromagnetics. Thus, time series analysis has become one of the staples for science and engineering technological fields.

16.4 TIME SERIES DATA ANALYSIS

Time series data refers to a sequence of data points collected or recorded over time. These data points are typically ordered chronologically and can be collected at regular or irregular intervals. Time series analysis involves studying the patterns, trends, and behaviors within the data to make predictions or gain insights into the underlying processes.

Key components of time series data include:

1. **Time Stamp:** Each data point is associated with a specific timestamp or time interval, indicating when the observation was made.
2. **Observations:** These are the values or measurements associated with each timestamp, representing the variable of interest being tracked over time.
3. **Seasonality:** Some time series data may exhibit regular patterns or cycles, known as seasonality, which could be daily, weekly, monthly, or annual.
4. **Trends:** Trends represent long-term movements or changes in the data over time. They can be ascending, descending, or remain relatively stable.
5. **Noise:** Random fluctuations or irregularities in the data that are not part of the underlying pattern or trend.

16.5 TECHNIQUES OF TIME SERIES ANALYSIS:

1. **Descriptive Analysis:** Examining the basic features of the data, such as mean, median, and standard deviation, to understand its characteristics.
2. **Visualization:** Plotting time series data on graphs, such as line charts, to visually identify patterns, trends, and anomalies.
3. **Smoothing:** Applying techniques like moving averages to reduce noise and highlight underlying trends.
4. **Decomposition:** Breaking down a time series into its constituent components, such as trend, seasonality, and residual (error).

5. **Forecasting:** Using statistical models or machine learning algorithms to predict future values based on historical data.

6. **Autocorrelation and Cross-Correlation:** Examining the correlation between a time series and its lagged values or the correlation between two different time series.

Time series analysis is widely used in various fields, including finance, economics, meteorology, signal processing, and many others. It helps in making informed decisions, detecting anomalies, and forecasting future trends based on historical data.

16.6 UTILITY OF TIME SERIES ANALYSIS

The analysis of Time Series is of great significance not only to the economist and businessman but also to the scientist, geologist, biologist, research worker, etc., for the reasons given below:

(1) It helps in Understanding Past Behaviours: By observing data over a period of time one can easily understanding what changes have taken place in the past, Such analysis will be extremely helpful in producing future behavior.

(2) It helps in Planning Operations : Plans for the future cannot be made without forecasting events and relationship they will have. Statistical techniques have been evolved which enable time series to be analyzed in such a way that the influences which have determined the form of that series to be analyzed in such a way that the influences which have determined the form of that series may be ascertained. If the regularity of occurrence of any feature over a sufficient long period could be clearly established then, within limits, prediction of probable future variations would become possible.

(3) It helps in Evaluating Current Accomplishments: The performance can be compared with the expected performance and the cause of variation analyzed. For example, if expected sale for 1995 was 10,000 refrigerators and the actual sale was only 9,000, one can investigate the

cause for the shortfall in achievement. Time series analysis will enable us to apply the scientific procedure of "holding other things constant" as we examine one variable at a time. For example, if we know how much the effect of seasonality on business is we may devise ways and means of ironing out the seasonal influence or decreasing it by producing commodities with complementary seasons.

(4) It Facilitates Comparison : Different time series are often compared and important conclusions drawn there from.

However, one should not be led to believe that by time series analysis one can foretell with 100percnet accuracy the course of future events. After all, statisticians are not foretellers. This could be possible only if the influence of the various forces which affect these series such as climate, customs and traditions, growth and decline factors and the complex forces which proclimate, customs and traditions, growth and decline factors and the complex forces which produce business cycles would have been regular in their operation. However, the facts of life reveal that this type of regularity does not exist. But this then does not mean that time series analysis is of value. When such analysis is couples with a careful examination of current business indicators once can undoubtedly improve substantially upon guest mates (i.e., estimates based upon pure guesswork) in forecasting future business conditions.

16.7 IMPORTANCE OF TIME SERIES

Some of the Importance are mentioned below:

- It is helpful for many organizations to forecast their business profit or loss trends. Thus essential business decisions can facilitate development.
- It is helpful to compare the present trend with the past trend that has already happened so the future trend can be estimated and prepared.
- The cycle variations over a period using time series will allow us to understand the business cycle quite effectively.
- It is helpful to understand the correlated seasonal trends of the data.

- It is also helpful in the quality control process to predicate the quality trend over time.
- Suppose we receive the complex signal pattern for it. In that case, we can apply transformations, such as Fourier analysis, to denoise the graph and break the complex pattern into simpler patterns. Hence, a better understanding can be achieved.
- It is also helpful to understand how an event can change its feature over a period of time. Hence, reliability, flexibility, and other essential features can be predicated.

16.8 WHY DO WE NEED TIME SERIES ANALYSIS?

Time series analysis is essential for several reasons in various fields. Here are some key reasons why it is important:

1. Understanding Trends and Patterns:

- Time series analysis helps in identifying and understanding trends and patterns within the data. This information is valuable for decision-making and strategic planning.

2. Forecasting Future Trends:

- By analyzing historical time series data, it is possible to build models that can predict future trends. This is crucial for businesses and organizations to make informed decisions and prepare for upcoming changes.

3. Anomaly Detection:

- Time series analysis can help detect abnormal or unexpected behavior in a dataset. Identifying anomalies is crucial in fields such as finance, where unusual patterns may indicate fraudulent activities, or in industrial settings, where anomalies might signal equipment malfunctions.

4. Resource Allocation:

- Understanding historical patterns allows for better resource allocation.

For example, businesses can optimize inventory management, workforce planning, and production schedules based on past trends.

5. **Financial Analysis:**
   - In finance, time series analysis is used for stock price prediction, risk management, and portfolio optimization. Traders and investors rely on historical data to make decisions about buying, selling, or holding financial assets.
6. **Economic Analysis:**
   - Governments and policymakers use time series analysis to monitor and analyze economic indicators such as GDP, unemployment rates, inflation, and interest rates. This information is critical for formulating economic policies.
7. **Demand Planning:**
   - Businesses use time series analysis to forecast demand for their products or services. This is particularly important in industries like retail, where inventory management and supply chain efficiency are key factors.
8. **Quality Control and Manufacturing:**
   - In manufacturing, time series analysis can be employed to monitor the quality of products and identify any deviations from the standard. This helps in maintaining consistency and improving production processes.
9. **Environmental Monitoring:**
   - Meteorological and environmental data are often represented as time series. Analyzing this data helps in understanding climate patterns, predicting natural disasters, and making decisions related to resource management and conservation.
10. **Healthcare and Medicine:**
    - In healthcare, time series analysis can be applied to monitor patient vital signs, disease progression, and the effectiveness of treatments. It aids in making informed decisions about patient care and resource allocation.

In summary, time series analysis provides valuable insights into the past, helps in forecasting future trends, and supports decision-making across various domains, contributing to improved efficiency, resource optimization, and strategic planning.

16.9 SUMMARY

In conclusion, time series analysis stands as a pivotal tool in extracting valuable insights from sequential data, offering a systematic approach to understanding and forecasting temporal patterns. The comprehensive examination of time series involves breaking down data into its fundamental components, including trend, seasonality, cyclic patterns, and irregular fluctuations. Each component contributes to a holistic view of the data's behavior over time.

Various approaches, such as time series models (e.g., ARIMA, SARIMA), regression analysis, VAR models, Bayesian econometrics, machine learning techniques, and ensemble methods, cater to different characteristics and complexities inherent in time series data. These approaches enable analysts and researchers to model and forecast economic variables, providing a lens through which to interpret historical trends and make informed predictions about the future.

As technology advances, the future of time series analysis holds promise. Integration with sophisticated machine learning methodologies, real-time data processing, and the development of more interpretable models are areas that continue to evolve, enhancing the precision and applicability of time series forecasting.

However, successful time series analysis requires careful consideration of data quality, thorough model validation, and a critical examination of underlying assumptions. Acknowledging these considerations ensures the reliability and relevance of the insights gained from the analysis.

In summary, time series analysis is a dynamic and adaptable methodology, playing a crucial role in diverse fields such as finance, economics, meteorology, and more. Its ability to capture and interpret temporal patterns

positions it as an indispensable tool for researchers, analysts, and decision-makers seeking to navigate and anticipate trends in sequential data.

16.10 GLOSSARY

- **Time Series Data Analysis:** Time series data analysis involves the exploration, modeling, and interpretation of data collected over time. Time series analysis is particularly useful for understanding temporal patterns, trends, and dependencies within the data
- **Forecasting Future Trends:** Forecasting future trends involves predicting the future direction or pattern of a variable, typically based on historical data, statistical models, or other analytical methods. It's a process of estimating what is likely to happen in the future, considering past patterns and current information. Forecasting is used in various fields, including business, finance, economics, meteorology, and many others, to make informed decisions and plans.
- **Techniques of Time Series Analysis:** Several commonly used techniques in time series analysis are employed to explore, model, and make predictions based on data collected over time. Here are some widely used techniques:

1. Descriptive Statistics:

- Basic statistical measures such as mean, median, standard deviation, and skewness provide an initial summary of the time series data, offering insights into its central tendency and variability.

2. Time Series Visualization:

- Visual inspection of time series data using line charts, scatter plots, and other graphical representations helps identify trends, seasonality, and potential outliers.

3. Decomposition:

- Decompose the time series into its constituent parts, typically trend,

seasonality, and residual components. This separation aids in understanding the underlying patterns and variations within the data.

4. **Autocorrelation and Partial Autocorrelation Functions (ACF and PACF):**

- ACF and PACF plots reveal the autocorrelation and partial autocorrelation at different lags, helping to identify the appropriate order of autoregressive (AR) and moving average (MA) terms in time series models.

5. **Stationarity Testing:**

- Conduct tests, such as the Augmented Dickey-Fuller (ADF) test or the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, to assess whether the time series is stationary. Stationarity is often a prerequisite for many time series models.

16.11 SELF ASSESSMENT QUESTIONS

1. **How is time series data analysis useful in Statistics?**

_____

_____

_____

2. **Explain the concept of econometric forecasting?**

_____

_____

_____

3. **Why do we need time series analysis?**

_____

_____

16.12 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics",4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell*. *Andrew W. Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*
10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"
11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"
12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"
13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Ruud, Paul A., 2000 “An Introduction to Classical Econometric Theory”, Oxford, 2000.

15. Wooldridge, J M, 2009 “Introductory Econometrics – A Modern Approach” (4th ed), South-Western, 2009.

****

UNIT-IV **LESSON NO. 17**

TIME SERIES ECONOMETRICS

COMPONENT OF TIME SERIES, SECULAR TREND, SEASONAL VARIATION, CYCLICAL VARIATION, PRELIMINARY ADJUSTMENT BEFORE ANALYSING TIME SERIES

STRUCTURE

17.1 INTRODUCTION

How do people get to know that the price of a commodity has increased over a period of time? They can do so by comparing the prices of the commodity

for a set of a time period. A set of observations ordered with respect to the successive time periods is a time series.

In other words, the arrangement of data in accordance with their time of occurrence is a time series. It is the chronological arrangement of data. Here, time is just a way in which one can relate the entire phenomenon to suitable reference points. Time can be hours, days, months or years.

A time series depicts the relationship between two variables. Time is one of those variables and the second is any quantitative variable. It is not necessary that the relationship always shows increment in the change of the variable with reference to time. The relation is not always decreasing too.

It may be increasing for some and decreasing for some points in time. Can you think of any such example? The temperature of a particular city in a particular week or a month is one of those examples.

17.2 COMPONENTS OF TIME SERIES

Time series data can be decomposed into several components, each of which captures a different aspect of the underlying patterns and variations. The primary components of a time series are often described using the decomposition framework, which breaks down the series into the following parts:

Time series data can be decomposed into several components, each contributing to the overall behavior of the series. The main components of a time series are:

Trend (T):

Definition: The long-term movement or general direction in the data.

Characteristics: It represents the overall pattern of growth, decline, or stability in the data over an extended period.

Example: An increasing trend in monthly sales data for a product.

Seasonality (S):

Definition: Repeating patterns or fluctuations in the data that occur at regular intervals.

Characteristics: Seasonal patterns often follow a fixed time frame, such as daily, weekly, monthly, or yearly cycles.

Example: Increased ice cream sales during summer months.

Cyclical Patterns (C):

Definition: Longer-term undulating patterns that do not have a fixed period like seasonality.

Characteristics: Cycles are less regular and may not occur at fixed intervals. They often represent economic or business cycles.

Example: Economic recessions and expansions.

Irregular or Residual (ε or R):

Definition: Unpredictable or random fluctuations in the data that cannot be attributed to trend, seasonality, or cyclic patterns.

Characteristics: Irregular components represent the noise or random variability in the data.

Example: Unforeseen events affecting sales, like a sudden product recall.

The time series itself (Yt) can be expressed as the sum of these components:

$$Yt = Tt + St + Ct + \varepsilon t$$

In this equation:

YT is the observed value at time t,

Tt is the trend component at time t,

St is the seasonal component at time t,

Ct is the cyclical component at time t, and

εt is the irregular or residual component at time t.

Analyzing and understanding these components is crucial for time series

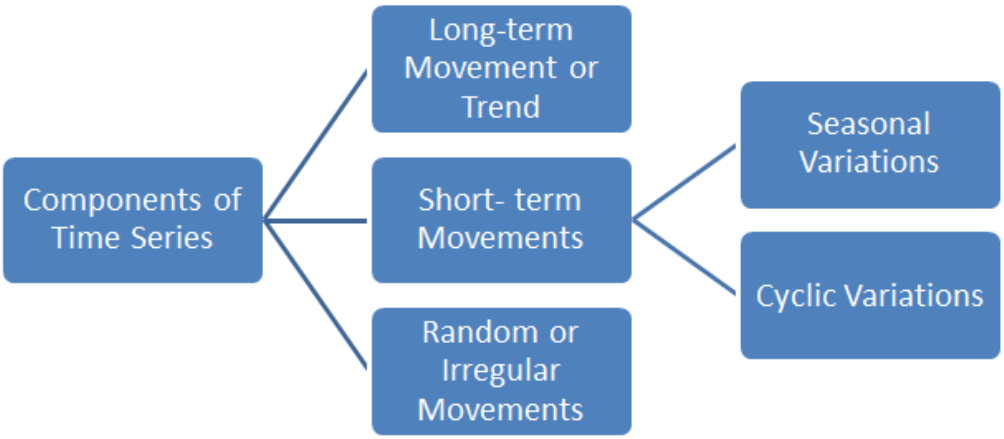
forecasting and modeling. Techniques such as decomposition methods (e.g., moving averages, exponential smoothing) help in isolating and estimating these components, allowing for more accurate predictions and insights into the underlying patterns of the time series data.

17.2.1 Components for Time Series Analysis

The various reasons or the forces which affect the values of an observation in a time series are the components of a time series. The four categories of the components of time series are

- Trend
- Seasonal Variations
- Cyclic Variations
- Random or Irregular movements

Seasonal and Cyclic Variations are the periodic changes or short-term fluctuations.



17.2.2 Trend

The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average

tendency. It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.

It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be upward, downward or stable. The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.

Linear and Non-Linear Trend

If we plot the time series values on a graph in accordance with time t. The pattern of the data clustering shows the type of trend. If the set of data cluster more or less round a straight line, then the trend is linear otherwise it is non-linear (Curvilinear).

17.2.3 Periodic Fluctuations

There are some components in a time series which tend to repeat themselves over a certain period of time. They act in a regular spasmodic manner.

Seasonal Variations

These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.

These variations come into play either because of the natural forces or man-made conventions. The various seasons or climatic conditions play an important role in seasonal variations. Such as production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.

The effect of man-made conventions such as some festivals, customs, habits, fashions, and some occasions like marriage is easily noticeable. They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Cyclic Variations

The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.

It is a four-phase cycle comprising of the phases of prosperity, recession, depression, and recovery. The cyclic variation may be regular are not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Random or Irregular Movements

There is another factor which causes the variation in the variable under study. They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

17.3 MATHEMATICAL MODEL FOR TIME SERIES ANALYSIS

Mathematical Model for Time Series Analysis

Mathematically, a time series is given as

$$yt = f(t)$$

Here, yt is the value of the variable under study at time t. If the population is the variable under study at the various time period t1, t2, t3, … , tn. Then the time series is

$t: t_1, t2, t_3, \ldots, t_n$

$y_t: y_{t1}, y_{t2}, y_{t3}, \ldots, y_{tn}$

or, $t: t_1, t_2, t_3, \ldots, t_n$

$y_t: y_1, y_2, y_3, \ldots, y_n$

Additive Model for Time Series Analysis

If y_t is the time series value at time t. T_t, S_t, C_t, and R_t are the trend value, seasonal, cyclic and random fluctuations at time t respectively. According to the Additive Model, a time series can be expressed as

$$y_t = T_t + S_t + C_t + R_t.$$

This model assumes that all four components of the time series act independently of each other.

Multiplicative Model for Time Series Analysis

The multiplicative model assumes that the various components in a time series operate proportionately to each other. According to this model

$$y_t = T_t \times S_t \times C_t \times R_t$$

Mixed models

Different assumptions lead to different combinations of additive and multiplicative models as

$$y_t = T_t + S_t + C_tR_t.$$

The time series analysis can also be done using the model $y_t = T_t + S_t \times C_t \times R_t$ or $y_t = T_t \times C_t + S_t \times R_t$ etc.

If *Y(t)* represents the observed time series at time *t*, the decomposition can be expressed as:

Y(t) =Trend+Seasonality+Cyclic+Irregular*Y(t)*

=Trend*(t)*+Seasonality*(t)*+Cyclic*(t)*+Irregular*(t)*

Importance of Decomposition:

Decomposing time series data into these components is crucial for better understanding the underlying structure and for building more accurate forecasting models. By isolating these components, analysts can make more informed decisions and predictions based on the inherent patterns within the data.

Decomposition Techniques:

Several techniques can be used for time series decomposition, including

moving averages, exponential smoothing, and more advanced methods like the Seasonal-Trend decomposition using LOESS (STL) or the Hodrick-Prescott filter.

Understanding and appropriately addressing these components are essential for effective time series analysis and forecasting. Different statistical methods and machine learning models can be applied based on the nature of these components to extract meaningful insights from time series data.

17.4 USES OF TIME SERIES

- The most important use of studying time series is that it helps us to predict the future behaviour of the variable based on past experience
- It is helpful for business planning as it helps in comparing the actual current performance with the expected one
- From time series, we get to study the past behaviour of the phenomenon or the variable under consideration
- We can compare the changes in the values of different variables at different times or places, etc.

17.5 IMPORTANCE OF TIME SERIES

In summary, the importance of understanding the components of time series lies in the ability to extract meaningful information, make informed predictions, and guide decision-making processes. Whether in business, economics, finance, or other fields, recognizing trends, seasonality, cyclic patterns, and irregularities provides a foundation for strategic planning, resource allocation, and proactive responses to changing conditions. Time series analysis becomes a powerful tool when these components are appropriately identified and incorporated into models and forecasts.

17.6 PRELIMINARY ADJUSTMENTS BEFORE ANALYSING TIME SERIES

Before conducting a time series analysis, it's important to perform some

preliminary adjustments and checks to ensure the data is suitable for analysis. Here are some common preliminary adjustments:

1. **Data Cleaning:**
   - Check for missing values and decide on a strategy for handling them (e.g., imputation or removal).
   - Identify and handle outliers or anomalies that may distort the analysis.
   - Ensure data is in a consistent format, and there are no errors or inconsistencies.
2. **Consistent Time Intervals:**
   - Ensure that the time series data has a consistent and regular time interval between observations. If not, consider resampling or interpolating to create regular intervals.
3. **Stationarity:**
   - Stationarity is an important assumption for many time series models. Check if the data is stationary, meaning that its statistical properties do not change over time.
   - If the data is not stationary, consider transforming it using techniques like differencing to stabilize the mean or variance.
4. **Seasonal Adjustment:**
   - Identify and, if necessary, remove seasonal effects from the data. This involves adjusting for regular patterns that repeat at known intervals.
   - Seasonal adjustment can improve the accuracy of models by isolating underlying trends and reducing noise.
5. **Transformations:**
   - Consider applying transformations to stabilize variance or make the data more linear. Common transformations include logarithmic or square root transformations.
   - Transformations can be useful when dealing with data that exhibits

non-constant variance or when the relationship between variables is not linear.

6. **Calendar Adjustments:**
   - If the time series data is influenced by calendar effects (e.g., weekends, holidays), consider making calendar adjustments to account for these variations.
7. **Data Scaling:**
   - Standardize or normalize the data if there are significant differences in scales between variables. This is important when using certain modeling techniques that are sensitive to scale.
8. **Visual Exploration:**
   - Plot the time series data to visually inspect for trends, seasonality, and other patterns.
   - Examine autocorrelation and partial autocorrelation plots to identify potential temporal dependencies.
9. **Check for Trends and Breakpoints:**
   - Look for trends or structural breaks in the time series that may impact the analysis. Structural breaks can occur due to changes in data collection methods or external events.
10. **Data Documentation:**
    - Keep detailed documentation about the data, including the source, collection methods, and any transformations applied. This information is essential for reproducibility.

By addressing these preliminary steps, you create a cleaner and more suitable dataset for time series analysis. This process helps ensure that the assumptions of the chosen time series model are met and that the analysis provides meaningful insights and accurate forecasts.

17.7 SUMMARY

Time series refers to a series of data points or observations recorded or measured sequentially over time. Analyzing time series data is crucial in various fields, including finance, economics, environmental science, signal processing, and many others. Here's a summary covering key aspects of time series:

1. Definition:

- A time series is a sequence of data points collected or recorded at successive, equally spaced intervals.

2. Components of Time Series:

- **Trend:** The long-term movement or direction in the data.
- **Seasonality:** Repeating patterns or fluctuations that occur at regular intervals.
- **Cyclic Patterns:** Longer-term undulating patterns that are not strictly periodic.
- **Irregularity/Noise:** Random fluctuations or noise in the data.

3. Time Series Analysis Techniques:

- **Descriptive Analysis:** Examining patterns, trends, and summary statistics.
- **Smoothing:** Removing noise to highlight underlying trends or patterns.
- **Decomposition:** Separating a time series into its constituent components.
- **Stationarity Testing:** Assessing whether statistical properties of a time series remain constant over time.
- **Modeling:** Using mathematical models (e.g., autoregressive integrated moving average - ARIMA, seasonal decomposition of time series - STL) to capture and predict patterns in the data.
- **Forecasting:** Predicting future values based on historical patterns.

4. **Challenges in Time Series Analysis:**
   - **Non-Stationarity:** Many time series exhibit changing statistical properties over time.
   - **Seasonal Adjustments:** Handling periodic variations that may affect analysis and forecasting.
   - **Data Gaps and Outliers:** Dealing with missing values and anomalous observations.
   - **Model Complexity:** Selecting appropriate models that balance accuracy and simplicity.
   - **Overfitting:** Ensuring models generalize well to new data.
5. **Applications:**
   - **Economics and Finance:** Stock prices, economic indicators, and financial market trends.
   - **Meteorology:** Climate patterns, temperature variations, and weather forecasts.
   - **Healthcare:** Patient monitoring, disease prevalence, and epidemic forecasting.
   - **Manufacturing:** Production and inventory tracking.
   - **Signal Processing:** Speech recognition, audio analysis, and communication systems.
6. **Time Series Data Visualization:**
   - **Line Charts:** Displaying trends over time.
   - **Seasonal Subseries Plots:** Highlighting seasonal patterns.
   - **Autocorrelation and Partial Autocorrelation Plots:** Assessing correlation between lagged observations.
7. **Software Tools:**
   - Various tools and programming languages like R, Python (with libraries

such as pandas, statsmodels), MATLAB, and specialized software like Tableau are commonly used for time series analysis.

Understanding and effectively analyzing time series data are essential for making informed decisions, predicting future trends, and gaining insights into the underlying processes represented by the data. Time series analysis provides valuable information for a wide range of applications, contributing to better planning, forecasting, and decision-making.

17.8 GLOSSARY

Components of Time Series: A time series is a series of data points ordered by time. It is a sequence of observations or measurements taken at successive or equally spaced points in time. The components of a time series can be decomposed into several elements, each contributing to the overall behavior of the series. The most commonly recognized components of a time series are:

1. **Trend:**
   - The long-term movement or general direction of the time series. It represents the underlying pattern in the data that shows whether the values are increasing, decreasing, or remaining relatively constant over time.
2. **Seasonality:**
   - Repeating patterns or cycles in the data that occur at regular intervals. Seasonal patterns often correspond to specific time periods, such as daily, weekly, monthly, or yearly cycles. For example, retail sales might have a seasonal pattern with increased sales during holiday seasons.
3. **Cyclic Component:**
   - Similar to seasonality, but the cycles in this component are not necessarily of fixed or repeating lengths. Cycles represent fluctuations that are not strictly tied to a specific time frame and may have varying durations.
4. **Irregular or Residual Component:**

- The random and unpredictable fluctuations in the data that cannot be attributed to the trend, seasonality, or cyclic components. This component represents the noise or residual errors in the time series.
- **Descriptive Analysis:** Examining patterns, trends, and summary statistics.
- **Smoothing:** Removing noise to highlight underlying trends or patterns.
- **Decomposition:** Separating a time series into its constituent components.
- **Stationarity Testing:** Assessing whether statistical properties of a time series remain constant over time.
- **Modeling:** Using mathematical models (e.g., autoregressive integrated moving average - ARIMA, seasonal decomposition of time series - STL) to capture and predict patterns in the data.
- **Forecasting:** Predicting future values based on historical patterns.
- **Non-Stationarity:** Many time series exhibit changing statistical properties over time.
- **Seasonal Adjustments:** Handling periodic variations that may affect analysis and forecasting.
- **Data Gaps and Outliers:** Dealing with missing values and anomalous observations.
- **Model Complexity:** Selecting appropriate models that balance accuracy and simplicity.
- **Overfitting:** Ensuring models generalize well to new data.

17.9 SELF ASSESSMENT QUESTIONS

1. What does the term ‘long period of time’ in a trend depict?

_____

_____

_____

2. For which type of data the seasonal fluctuations do not appear in a time series?

_____

_____

_____

3. How is time series data important in Econometrics ?

_____

_____

_____

17.10 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics",4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell*. *Andrew W. Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29, 1996*

10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"

11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"

12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"

13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Ruud, Paul A., 2000 "An Introduction to Classical Econometric Theory", Oxford, 2000.

15. Wooldridge, J M, 2009 "Introductory Econometrics – A Modern Approach" (4th ed), South-Western, 2009.

****

UNIT-IV **LESSON NO. 18**

TIME SERIES ECONOMETRICS

TIME SERIES ECONOMETRICS, STOCHASTIC PROCESSES, STATIONERY STOCHASTIC PROCESS, NON STATIONERY STOCHASTIC PROCESS

STRUCTURE

18.1 INTRODUCTION

Stochastic processes play a crucial role in the analysis of time series data. Time series data represents a sequence of observations collected over time, and stochastic processes provide a framework for modeling the uncertainty or randomness inherent in these observations. Here are some key concepts related to stochastic processes in the context of time series:

1. Stochastic Process:

- A stochastic process is a collection of random variables indexed by time. It represents the evolution of a system over time where randomness is involved.
- In time series analysis, the observed data is often assumed to be generated by some underlying stochastic process.

2. Stationarity:

- Stationarity is a key assumption in time series analysis. A time series is said to be stationary if its statistical properties, such as mean and variance, do not change over time.
- Weak stationarity requires constant mean and variance, while strong stationarity additionally requires that the joint distribution of any set of observations is the same for all time points.

3. Autoregressive (AR) Processes:

- An autoregressive process is a stochastic process where each value in the time series is a linear combination of its past values, plus a random error term.
- AR(p) models depend on the past p observations.

4. Moving Average (MA) Processes:

- A moving average process is a stochastic process where each value is a linear combination of current and past random error terms.
- MA(q) models depend on the past q random errors.

5. Autoregressive Integrated Moving Average (ARIMA) Models:

- ARIMA combines autoregressive, moving average, and differencing components to model different aspects of a time series.
- ARIMA(p, d, q) consists of autoregressive order p, differencing order d, and moving average order q.

6. Seasonal Time Series:

- Some time series exhibit periodic patterns known as seasonality. Seasonal processes involve repeating patterns over specific time intervals.
- Seasonal decomposition of time series can help separate the trend, seasonal, and residual components.

7. White Noise:

- White noise is a special case of a stochastic process where each observation is independent and identically distributed with a constant mean and variance.
- It serves as a baseline for comparing the randomness in a time series.

8. Estimation and Forecasting:

- Parameters of stochastic processes are often estimated from observed data using methods like maximum likelihood estimation (MLE).
- Forecasting involves predicting future values of a time series based on the estimated model parameters.

Understanding and modeling stochastic processes in time series analysis is essential for making predictions, detecting patterns, and gaining insights into the underlying dynamics of the observed data. Various statistical and machine learning techniques are employed for modeling and forecasting time series data in practice.

18.2 OBJECTIVES

Stochastic processes are employed in the analysis of time series data to achieve several objectives. These objectives help researchers, analysts, and practitioners understand and model the inherent randomness and dynamics present in the observed data. Here are some key objectives of using stochastic processes in time series analysis:

1. Modeling Randomness:

- One of the primary objectives is to model the random or stochastic nature of time series data. Stochastic processes provide a mathematical framework to describe the uncertainty and variability in observed sequences of data points.

2. Forecasting and Prediction:

- Stochastic processes are used to develop models that can forecast and predict future values of a time series. By capturing the underlying patterns and randomness, these models assist in making informed predictions about future observations.

3. Risk Assessment:

- In financial and economic time series, understanding and quantifying risk are crucial. Stochastic processes help in modeling and assessing the risk associated with various financial instruments, asset prices, and economic indicators.

4. Statistical Inference:

- Stochastic processes facilitate statistical inference by providing a probabilistic framework for hypothesis testing, confidence interval estimation, and assessing the significance of observed patterns or trends in time series data.

5. Anomaly Detection:

- Stochastic models can be used to identify anomalies or unusual patterns in time series data. Sudden deviations from the expected behavior can be

detected, which is important in various applications such as fraud detection or system monitoring.

6. **Dynamic Systems Modeling:**

- Stochastic processes help in modeling dynamic systems where the evolution of a variable over time is influenced by both deterministic trends and random shocks. This is particularly relevant in fields like physics, biology, and engineering.

7. **Parameter Estimation:**

- Stochastic models often involve parameters that need to be estimated from observed data. Techniques such as maximum likelihood estimation (MLE) are used to find the parameter values that best explain the observed time series.

8. **Seasonal and Trend Decomposition:**

- Stochastic processes assist in decomposing time series into components such as trend, seasonality, and residual randomness. This decomposition aids in understanding the underlying structure of the data and isolating specific patterns.

9. **Simulation and Monte Carlo Methods:**

- Stochastic processes enable the simulation of multiple possible future scenarios. Monte Carlo methods, based on stochastic simulations, can be used for risk analysis, option pricing, and other decision-making processes.

10. **Improving Decision-Making:**

- By providing a probabilistic framework and capturing the uncertainty inherent in time series data, stochastic processes contribute to more informed decision-making in various fields, including finance, economics, engineering, and environmental science.

In summary, the objectives of using stochastic processes in time series analysis are diverse, ranging from modeling randomness and forecasting future values to making statistical inferences and improving decision-making in a wide range of applications.

18.3 TYPES OF STOCHASTIC PROCESSES IN TIME SERIES

Various types of stochastic processes are used to model time series data, each with its own characteristics and applications. Here are some common types of stochastic processes used in time series analysis:

1. White Noise:

- White noise is a simple and fundamental stochastic process where each observation is independently and identically distributed with constant mean and variance. It serves as a baseline for randomness.

2. Random Walk:

- A random walk is a stochastic process where each value is determined by adding a random shock to the previous value. It is a simple model for trends in time series data.

3. Autoregressive (AR) Process:

- In an autoregressive process of order p (AR(p)), each value in the time series is a linear combination of its past p values plus a random error term. AR processes capture serial correlation in the data.

4. Moving Average (MA) Process:

- In a moving average process of order q (MA(q)), each value is a linear combination of current and past random error terms. MA processes are used to model short-term shocks.

5. Autoregressive Integrated Moving Average (ARIMA) Process:

- ARIMA combines autoregressive, differencing, and moving average components. It is represented as ARIMA (p, d, q), where p is the autoregressive order, d is the differencing order, and q is the moving average order.

6. Seasonal Processes:

- Seasonal processes involve periodic patterns that repeat over specific time intervals. Seasonal components are added to time series models to capture these recurring patterns.

7. **GARCH (Generalized Autoregressive Conditional Heteroskedasticity) Process:**
   - GARCH models are used to capture volatility clustering in financial time series. They model the conditional variance of the process, allowing for time-varying volatility.
8. **State Space Models:**
   - State space models are a general framework that combines both observed and unobserved components. They are useful for modeling complex time series data and are commonly used in econometrics.
9. **Long Memory Processes (Fractional Brownian Motion):**
   - Long memory processes exhibit persistent dependence over time, and fractional Brownian motion is a well-known example. These processes are characterized by slowly decaying autocorrelations.
10. **Markov Chains:**
    - Markov chains model a sequence of events where the probability of transitioning to a particular state depends only on the current state. They are used in applications such as queueing theory and simulation studies.
11. **Cointegrated Processes:**
    - Cointegrated processes involve multiple time series that share a common stochastic trend. They are often used in modeling relationships between non-stationary variables.
12. **Hidden Markov Models (HMM):**
    - HMMs are used when the underlying state of a system is not directly observable. They involve observable emissions influenced by an unobservable Markov process.

These are just a few examples of the many stochastic processes used in time series analysis. The choice of a particular model depends on the characteristics of the data and the specific objectives of the analysis. Researchers

often use a combination of these models to capture various aspects of the underlying processes in time series data.

18.4 STATIONARY STOCHASTIC PROCESSES IN TIME SERIES

Stationary stochastic processes play a crucial role in time series analysis, as the assumption of stationarity simplifies the modeling and analysis of the data. A stationary process is one in which the statistical properties of the data do not change over time. There are different forms of stationarity, and each has its own implications for the modeling and interpretation of time series data. Here are some key concepts related to stationary stochastic processes in time series:

1. Strict (Strong) Stationarity:

- A stochastic process is strictly stationary if the joint distribution of any set of observations is the same for all time points.
- Under strict stationarity, the moments of the distribution (mean, variance, etc.) are constant over time.

2. Weak (Second-Order) Stationarity:

- A weaker form of stationarity is weak stationarity, where the mean and variance of the process are constant over time, and the covariance between any two observations depends only on the time lag between them.
- Weak stationarity requires that the first and second moments of the process are time-invariant.

3. Covariance Stationarity:

- A process is covariance stationary if its mean and variance are constant over time, and the autocovariance function (ACF) depends only on the time lag.
- The autocorrelation function (ACF) is also constant over time under covariance stationarity.

4. Trend-Stationary and Difference-Stationary Processes:

- In the context of non-stationary processes, a time series can be decomposed into a trend component and a stationary component. If the detrended series is stationary, it is difference-stationary. If both the trend and the series are stationary, it is trend-stationary.

5. Unit Root and Non-Stationarity:

- A unit root is a characteristic of non-stationary time series. If a time series has a unit root, it implies that the series is not stationary in its levels.
- Differencing is often applied to make non-stationary series stationary by removing trends or seasonality.

6. Implications for Modeling:

- Stationarity simplifies the modeling process, as the statistical properties of the process do not change over time. This stability makes it easier to estimate parameters and make predictions.
- Many time series models, such as autoregressive (AR) and moving average (MA) models, assume stationarity for their validity.

7. Testing for Stationarity:

- Various statistical tests, such as the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, are used to assess stationarity in time series data.

8. Importance in Forecasting:

- Stationarity is essential for accurate forecasting. Models built on stationary data are more likely to generalize well to future observations.

It's important to note that not all time series data are stationary, and transformations or differencing may be required to achieve stationarity. The choice of the appropriate stationarity assumption depends on the characteristics of the data and the specific requirements of the analysis.

18.5 IMPORTANCE OF STATIONARY STOCHASTIC PROCESSES

The assumption of stationarity in stochastic processes is crucial in time series analysis for several reasons. Stationarity simplifies the modeling process, enhances the interpretability of statistical properties, and ensures the reliability of various statistical techniques. Here are some key reasons highlighting the importance of stationary stochastic processes in time series:

1. Simplified Modeling:

- Stationarity simplifies the modeling process by making the statistical properties of the time series constant over time. This stability facilitates the use of simpler and more interpretable models.

2. Parameter Estimation:

- In stationary processes, the parameters of the model remain constant over time, making parameter estimation more reliable. This enhances the accuracy of parameter estimates and improves the interpretability of the model.

3. Autoregressive and Moving Average Models:

- Many commonly used time series models, such as autoregressive (AR) and moving average (MA) models, assume stationarity for their validity. Stationary processes are essential for accurately capturing and modeling the temporal dependencies in the data.

4. Forecasting Accuracy:

- Stationary time series are generally easier to forecast accurately. Models built on stationary data are more likely to provide reliable predictions for future observations, contributing to the effectiveness of forecasting.

5. Statistical Tests:

- Various statistical tests, such as the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, are designed to assess stationarity in time series data. These tests help analysts

determine whether differencing or transformations are needed to achieve stationarity.

6. **Mean and Variance Stability:**

- Stationary processes have constant mean and variance over time. This stability simplifies the interpretation of statistical properties and ensures that the central tendencies and dispersion of the data do not change systematically.

7. **Autocorrelation Function (ACF):**

- Stationarity implies that the autocorrelation function (ACF) remains constant over time. A stable ACF simplifies the analysis of temporal dependencies, making it easier to identify and model patterns in the data.

8. **Avoidance of Spurious Relationships:**

- Non-stationary data may exhibit spurious correlations and trends. Stationarity helps avoid misinterpretation of relationships between variables, ensuring that observed patterns are not artifacts of changing statistical properties.

9. **Stable Long-Term Behavior:**

- Stationary processes exhibit stable long-term behavior, which is important for understanding and modeling the underlying dynamics of the system. This stability contributes to the robustness of the model over extended time periods.

10. **Facilitation of Time-Invariant Assumptions:**

- Stationarity supports the assumption that the properties of the time series are time-invariant. This is crucial for the validity of many statistical techniques and models.

While stationarity is a desirable property, it's important to note that not all time series data are stationary. In cases where non-stationarity is observed, techniques such as differencing or transformations may be applied to achieve stationarity, allowing for the use of stationary models. The choice of the

appropriate stationarity assumption depends on the characteristics of the data and the specific objectives of the analysis.

18.6 TYPES OF STATIONARY STOCHASTIC PROCESSES IN TIME SERIES

Stationary stochastic processes in time series analysis can take different forms, and the type of stationarity depends on the specific characteristics of the process. Here are several types of stationary stochastic processes commonly encountered in time series analysis:

1. Strict (Strong) Stationarity:

- A stochastic process is strictly stationary if the joint distribution of any set of observations is the same for all time points. It implies that all moments of the distribution, including mean, variance, and higher-order moments, are constant over time.

2. Weak (Second-Order) Stationarity:

- Weak stationarity is a less restrictive form that requires the first and second moments of the process to be constant over time. This implies a constant mean and variance, and the covariance between any two observations depends only on the time lag.

3. Covariance Stationarity:

- Covariance stationarity is a specific form of weak stationarity. It implies that the mean and variance are constant over time, and the autocovariance function (ACF) depends only on the time lag. This is a common form of stationarity assumed in many time series models.

4. Trend-Stationary Process:

- A time series is trend-stationary if it can be decomposed into a deterministic trend component and a stationary component. The stationary component remains constant over time, satisfying the conditions of weak stationarity.

5. Difference-Stationary Process:

- A time series is difference-stationary if the differenced series (the series of first differences) is stationary. This is a common approach to achieving stationarity, especially for non-stationary time series with trends or seasonality.

6. Seasonal Stationarity:

- Seasonal stationarity refers to a time series that exhibits constant statistical properties within each season. Seasonal adjustments are often applied to make the data stationary and remove recurring patterns.

7. Cointegrated Processes:

- Cointegration involves multiple time series that share a common stochastic trend. When the linear combination of these non-stationary variables yields a stationary series, they are said to be cointegrated.

8. Homogeneous Markov Chains:

- In a homogeneous Markov chain, the transition probabilities between states remain constant over time. This implies that the process is stationary with respect to its transition probabilities.

9. Circular Stationarity:

- Circular stationarity is applicable to processes involving circular data, such as angles or directions. It ensures that the distribution of the circular data remains constant over time.

10. Mixing Stationarity:

- A process is mixing stationary if the dependence between distant observations diminishes as the time lag increases. This form of stationarity is useful in analyzing the long-term behavior of a process.

It's important to note that the choice of stationarity depends on the specific characteristics of the time series data and the requirements of the analysis. In practice, statistical tests, such as the Augmented Dickey-Fuller (ADF) test or the

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, are often used to assess stationarity and guide the selection of an appropriate stationary model.

18.7 NON STATIONARY STOCHASTIC PROCESSES IN TIME SERIES

Non-stationary stochastic processes in time series analysis are those processes whose statistical properties change over time. The presence of non-stationarity can complicate the modeling and analysis of time series data. Here are some common types of non-stationary stochastic processes:

1. Trend:

- A time series exhibits a trend if there is a systematic upward or downward movement over time. The mean of the series is not constant, and the statistical properties change with time. Linear trends, quadratic trends, and exponential trends are common examples.

2. Seasonality:

- Seasonality refers to regular and predictable patterns that repeat at fixed intervals. These patterns may be daily, monthly, or yearly, and they introduce non-stationarity in the form of periodic fluctuations.

3. Cyclical Components:

- Cyclical components represent longer-term periodic patterns that are not strictly tied to fixed intervals. These cycles can result from economic fluctuations, business cycles, or other recurring but non-seasonal patterns.

4. Deterministic Components:

- Deterministic components introduce non-stationarity through fixed, predictable patterns that are not due to randomness. These may include step functions, ramps, or other deterministic trends.

5. Explosive Processes:

- Processes with trends that exhibit explosive behavior, such as exponential growth or decline, lead to non-stationarity. This often requires

transformations or differencing to stabilize the variance or make the series stationary.

6. **Unit Root Processes:**
   - A unit root in a time series indicates that the series is non-stationary. Unit root processes, such as random walks, have a long-term tendency to move away from the mean, making them non-stationary in levels.
7. **Integrated Processes (I(d)):**
   - Integrated processes involve differencing the time series data to achieve stationarity. A time series is said to be integrated of order d (I(d)) if it requires d differences to become stationary.
8. **Changing Variance:**
   - Heteroscedasticity, or changing variance over time, can introduce non-stationarity. Volatility clustering, where periods of high volatility are followed by periods of low volatility, is an example commonly observed in financial time series.
9. **Non-constant Autocorrelations:**
   - Non-stationary processes may exhibit autocorrelations that change over time. This lack of stability in the autocorrelation structure complicates the modeling and forecasting of the time series.
10. **Structural Breaks:**
    - Structural breaks occur when the underlying data-generating process changes abruptly at a certain point in time. These breaks introduce non-stationarity and may require modeling the data in different segments.

Addressing non-stationarity is essential for accurate modeling and forecasting. Common approaches to handling non-stationary time series data include differencing, transforming, or detrending the series to achieve stationarity. Unit root tests, such as the Augmented Dickey-Fuller (ADF) test, are often used to detect the presence of non-stationarity in a time series. Once non-stationarity is identified, appropriate transformations or differencing can be applied to make the data suitable for stationary time series models.

18.8 TYPES OF NON STATIONARY STOCHASTIC PROCESSES IN TIME SERIES

Non-stationary stochastic processes in time series analysis exhibit statistical properties that change over time. Non-stationarity can take various forms, and understanding these types is crucial for appropriate modeling and analysis. Here are some common types of non-stationary stochastic processes:

1. Trend Non-Stationarity:

- Time series with a systematic and persistent upward or downward movement over time exhibit trend non-stationarity. The mean of the series changes, violating the stationarity assumption.

2. Seasonal Non-Stationarity:

- Seasonal non-stationarity arises when a time series exhibits regular patterns or cycles that repeat at fixed intervals. Seasonal fluctuations can introduce non-stationarity in the form of periodic variations.

3. Cyclical Non-Stationarity:

- Cyclical non-stationarity involves longer-term patterns or cycles that are not strictly tied to fixed intervals. These cycles may represent economic fluctuations, business cycles, or other recurring but non-seasonal patterns.

4. Deterministic Non-Stationarity:

- Deterministic non-stationarity occurs when there are predictable and fixed patterns in the time series that are not due to random variation. This includes step functions, ramps, or other deterministic trends.

5. Explosive Processes:

- Time series with explosive behavior, such as exponential growth or decline, are non-stationary. These processes may require transformations or differencing to stabilize the variance or make the series stationary.

6. Unit Root Processes:

- A unit root in a time series indicates non-stationarity. Processes with unit

roots, like random walks, have a long-term tendency to move away from the mean, making them non-stationary in levels.

7. Integrated Processes (I(d)):

- Integrated processes involve differencing the time series data to achieve stationarity. A time series is said to be integrated of order d (I(d)) if it requires d differences to become stationary.

8. Changing Variance (Heteroscedasticity):

- Heteroscedasticity introduces non-stationarity through changing variance over time. This can manifest as volatility clustering, where periods of high volatility are followed by periods of low volatility.

9. Non-constant Autocorrelations:

- Time series with autocorrelations that change over time exhibit non-stationarity. The instability in the autocorrelation structure complicates modeling and forecasting.

10. Structural Breaks:

- Structural breaks occur when the underlying data-generating process changes abruptly at a certain point in time. These breaks introduce non-stationarity and may require modeling the data in different segments.

Addressing non-stationarity is essential for accurate modeling and forecasting in time series analysis. Common techniques for handling non-stationary data include differencing, transforming, or detrending the series to achieve stationarity. Unit root tests, such as the Augmented Dickey-Fuller (ADF) test, are commonly used to detect the presence of non-stationarity in a time series. Once identified, appropriate transformations or differencing methods can be applied to make the data suitable for stationary time series models.

18.9 IMPORTANCE OF NON STATIONARY STOCHASTIC PROCESSES IN TIME SERIES

Non-stationary stochastic processes in time series analysis are important to recognize and understand because they reflect dynamic and evolving patterns

in data. While stationarity simplifies modeling, the presence of non-stationarity introduces challenges and opportunities for richer analyses. Here are several reasons highlighting the importance of non-stationary stochastic processes in time series:

1. **Realism in Modeling:**
   - Many real-world phenomena exhibit non-stationary behavior. Ignoring non-stationarity may lead to oversimplified models that do not capture the complexities of the underlying processes. Acknowledging and modeling non-stationarity is essential for more realistic representations.
2. **Trend Detection and Economic Analysis:**
   - Non-stationary processes with trends can provide insights into economic and business trends. Analyzing trends in economic indicators, such as GDP or employment, allows for better understanding of long-term patterns and policy implications.
3. **Seasonal Patterns:**
   - Non-stationarity due to seasonality is common in various fields. Recognizing and modeling seasonal patterns in time series data are crucial for businesses, agriculture, retail, and other industries that are affected by recurring seasonal trends.
4. **Cyclical Fluctuations:**
   - Processes exhibiting cyclical non-stationarity can be valuable in understanding long-term economic cycles. Identifying and analyzing cyclical patterns is essential for making informed decisions in financial markets and economic policy.
5. **Structural Changes:**
   - Non-stationarity can be indicative of structural changes in the underlying data-generating process. Detecting these changes is important for understanding shifts in market conditions, consumer behavior, or other factors influencing the data.

6. **Volatility Clustering:**

- Heteroscedasticity and changing variances in non-stationary processes can capture periods of high and low volatility. Understanding volatility clustering is crucial in financial markets for risk management, option pricing, and portfolio optimization.

7. **Model Improvement:**

- Modeling non-stationary processes often requires advanced statistical methods. Addressing non-stationarity through techniques like differencing, detrending, or time-varying models can lead to improved model accuracy and better predictions.

8. **Dynamic Forecasting:**

- Non-stationary time series are dynamic and can exhibit evolving patterns. Recognizing non-stationarity allows for dynamic forecasting, where models adapt to changing conditions, providing more accurate predictions in dynamic environments.

9. **Long-Term Planning:**

- Non-stationary processes, especially those with trends and cycles, are crucial for long-term planning. Businesses, governments, and policymakers need to account for evolving patterns when making decisions that span extended time frames.

10. **Understanding System Dynamics:**

- Non-stationary processes provide insights into the dynamic nature of systems. Analyzing changes in statistical properties over time can help researchers understand the underlying dynamics of complex systems.

In summary, while stationarity simplifies modeling, non-stationary stochastic processes offer valuable information about the dynamic nature of time series data. Recognizing and appropriately addressing non-stationarity is essential for accurate modeling, forecasting, and gaining meaningful insights into the underlying processes driving the data.

18.10 SUMMARY

Time series econometrics is a branch of econometrics that focuses on the analysis, modeling, and forecasting of time-ordered data. It involves applying statistical and econometric techniques to study economic and financial variables over time. Here's a summary of key concepts and methods in time series econometrics:

1. Time Series Data:

- Time series data consists of observations on a variable or multiple variables collected or recorded over sequential points in time. It could be daily, monthly, quarterly, or yearly, depending on the frequency of data collection.

2. Stationarity:

- Stationarity is a fundamental concept in time series econometrics. Stationary processes have statistical properties that do not change over time. Stationarity simplifies modeling and is often assumed in many time series models.

3. Autocorrelation and Autocovariance:

- Autocorrelation measures the linear dependence between a variable's current value and its past values. Autocovariance is a related concept that quantifies the covariance between a variable's current and past values.

4. Autoregressive (AR) and Moving Average (MA) Models:

- AR models represent a variable as a linear combination of its past values, while MA models represent a variable as a linear combination of past white noise (random error) terms. Autoregressive Integrated Moving Average (ARIMA) models combine AR and MA components.

5. Seasonal Time Series Models:

- Seasonal patterns in time series data can be modeled using seasonal adjustments or through models like Seasonal Autoregressive Integrated Moving Average (SARIMA).

6. **Cointegration:**

- Cointegration deals with the long-run relationship between multiple time series. If variables are cointegrated, they share a common stochastic trend, which has important implications for modeling.

7. **Unit Root and Stationarity Tests:**

- Unit root tests, such as the Augmented Dickey-Fuller (ADF) test, help determine whether a time series is non-stationary. Stationarity tests, like the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test, provide complementary information.

8. **Vector Autoregression (VAR) Models:**

- VAR models extend the concept of autoregression to multiple variables, capturing the dynamic interactions between them over time.

9. **Granger Causality:**

- Granger causality tests whether past values of one time series variable help predict another variable. It is commonly used to assess causal relationships between economic variables.

10. **ARCH and GARCH Models:**

- ARCH (Autoregressive Conditional Heteroskedasticity) and GARCH models are used to model volatility clustering and changing variances in financial time series data.

11. **State Space Models:**

- State space models are a flexible framework that combines observed data with unobserved state variables. They are widely used in time series econometrics for structural modeling and forecasting.

12. **Panel Data Time Series Models:**

- Panel data, which involves observations on multiple entities over time, can be analyzed using panel data time series models like fixed effects or random effects models.

13. Dynamic Stochastic General Equilibrium (DSGE) Models:

- DSGE models are used in macroeconomics to study the dynamics of economies over time. They incorporate various economic agents and shocks to capture the interactions within an economy.

In summary, time series econometrics provides a toolkit for analyzing and modeling economic and financial data over time. It involves a range of statistical techniques to understand the dynamics, relationships, and trends within time-ordered datasets, making it a crucial field in empirical economics and finance.**Top of Form**

18.11 GLOSSARY

- **Non stationary stochastic processes:** A non-stationary stochastic process is a type of random process whose statistical properties, such as mean and variance, change over time. In contrast, a stationary stochastic process maintains constant statistical properties throughout its entire duration. Non-stationary processes are common in many real-world phenomena where underlying factors evolve or change with time.
- **Stationary stochastic processes:**A stationary stochastic process, also known simply as a stationary process, is a type of random process whose statistical properties do not change with time. The concept of stationarity is crucial in time series analysis as it simplifies the modeling and analysis of data. There are two main types of stationarity: strict stationarity and weak stationarity.
- **Strict Stationarity:** A stochastic process is strictly stationary if the joint probability distribution of any set of time points is invariant under time shifts. In other words, the entire probability distribution function (pdf) of the process remains constant over time. This is a very stringent condition and is rarely assumed in practice.
- **Weak Stationarity (Second-order Stationarity):** Weak stationarity is a more relaxed condition compared to strict stationarity. A stochastic process is weakly stationary if the mean, variance, and autocorrelation structure

(up to the second order) remain constant over time. This means that the expected value (mean) of the process is constant, the variance is constant, and the autocovariance between any two time points depends only on the time difference between them.

- **Trend:** A time series exhibits a trend if there is a systematic upward or downward movement over time. The mean of the series is not constant, and the statistical properties change with time. Linear trends, quadratic trends, and exponential trends are common examples.
- **Cyclical Components:**Cyclical components represent longer-term periodic patterns that are not strictly tied to fixed intervals. These cycles can result from economic fluctuations, business cycles, or other recurring but non-seasonal patterns.

18.12 SELF ASSESSMENT QUESTIONS

1. What are the Types of stochastic processes in time series ?

_____

_____

_____

2. Explain the Non stationary stochastic processes in time series?

_____

_____

_____

3. How is stationary stochastic processes important in Econometrics?

_____

_____

_____

18.13 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics", 4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell*. *Andrew W. Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*
10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"
11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"
12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"
13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Ruud, Paul A., 2000 “An Introduction to Classical Econometric Theory”, Oxford, 2000.

15. Wooldridge, J M, 2009 “Introductory Econometrics – A Modern Approach” (4th ed), South-Western, 2009.

****

UNIT-IV **LESSON NO. 19**

TIME SERIES ECONOMICTRICS

RANDOM WALK MODELS: COINTEGRATION, DETERMINISTIC AND STOCHASTIC TRENDS, UNIT ROOT TESTS

STRUCTURE

19.1 INTRODUCTION

Random walk models are mathematical models used to describe the stochastic or random movement of a variable over time. The term "random walk" reflects the idea that the variable's future values are determined by a series of random steps. There are different types of random walk models, and they are commonly employed in various fields, including finance, physics, and biology. Here are two primary types of random walk models:

1. Discrete Random Walk:

- In a discrete random walk, the variable changes its value at discrete time intervals.
- At each time step, the variable can move randomly in one of several directions, typically with equal probability.
- The simplest example is the one-dimensional random walk on a number line, where the variable can move one step to the left or right with equal probability.

2. Continuous Random Walk:

- In a continuous random walk, the variable changes continuously over time.
- The continuous-time analogue of the one-dimensional discrete random walk is often modeled using stochastic differential equations.
- Brownian motion, a specific type of continuous random walk, is widely used in physics and finance. It describes the random movement of particles suspended in a fluid.

Key Characteristics:

- **Markov Property:** Random walk models often exhibit the Markov property, meaning that the future values of the variable depend only on its current state and are independent of its past states.
- **No Memory:** Each step in a random walk is independent of previous

steps. This lack of memory is a crucial characteristic of random walk models.

- **Diffusion-Like Behavior:** Random walks often result in a diffusion-like behavior, where the variable's values spread out over time.

Applications:

1. **Financial Markets:** Random walk models are frequently used to model stock prices, assuming that price changes are unpredictable and follow a random pattern.
2. **Physics:** Brownian motion, a continuous random walk, is used to model the random movement of particles in a fluid. It has applications in fields such as statistical physics.
3. **Biology:** Random walk models are employed to describe various biological processes, such as the movement of cells or organisms in response to random stimuli.

While random walk models are useful for certain applications, they are also simplifications that may not capture all aspects of complex systems. Nevertheless, they provide valuable insights into the behavior of systems subject to random fluctuations.

19.2 WHAT IS COINTEGRATION

Cointegration is a statistical property that is often applied to time series data. It refers to a long-term relationship between two or more non-stationary time series variables. In a cointegrated relationship, while the individual variables may not be stationary (meaning their statistical properties change over time), a linear combination of them is stationary.

Here are the key concepts related to cointegration:

1. Non-stationary Time Series:

- A stationary time series has constant statistical properties over time, such as constant mean and variance.

- Non-stationary time series, on the other hand, exhibit trends, seasonality, or other patterns that change over time.

2. Cointegration:

- If two or more non-stationary time series have a stable long-term relationship, they are said to be cointegrated.
- Cointegration implies that even though the individual series may deviate from their mean values, a linear combination of them stays relatively constant over time.

3. Error Correction Mechanism (ECM):

- Cointegrated series often have an associated error correction mechanism. This mechanism helps to adjust for any short-term deviations from the long-term equilibrium relationship.
- In a cointegrated relationship, if the variables temporarily deviate from their long-term equilibrium, there is a force that pushes them back towards that equilibrium.

4. Example:

- A common example of cointegration is in financial markets, where the prices of two stocks might be non-stationary on their own. However, a linear combination of their prices (e.g., the spread between the two) may be stationary.

Applications:

1. Econometrics and Finance:

- Cointegration is widely used in econometrics and finance to model long-term relationships between economic variables, such as interest rates, exchange rates, or stock prices.

2. Macroeconomics:

- In macroeconomics, cointegration is employed to analyze relationships between economic indicators, like the relationship between income and consumption.

3. Time Series Analysis:

- Cointegration is a valuable tool in time series analysis for dealing with non-stationary data and understanding the underlying long-term relationships between variables.

Understanding cointegration is crucial for accurate modeling, forecasting, and making meaningful inferences when dealing with non-stationary time series data. The concept was introduced by Clive Granger and Robert Engle, who were awarded the Nobel Prize in Economic Sciences in 2003 for their work on cointegration and its implications for time series analysis.

19.3 OBJECTIVES OF COINTEGRATION

The main objectives of cointegration in time series analysis are to address issues related to non-stationary data and to capture the long-term relationships between variables. Here are the primary objectives of employing cointegration:

1. Handling Non-stationarity:

Cointegration helps address the problem of non-stationarity in time series data. Many economic and financial variables exhibit trends or other patterns that violate the assumptions of stationarity, making standard statistical methods less reliable. Cointegration allows for the analysis of relationships between non-stationary variables without the risk of spurious results.

2. Capturing Long-Term Relationships:

Cointegration is particularly useful when examining the long-term relationships between variables. It identifies whether there exists a stable combination of non-stationary variables that moves together over time. This is crucial for understanding the underlying economic or financial forces that link these variables in the long run.

3. Error Correction Modeling:

Cointegration often involves the use of error correction models (ECM). These models incorporate short-term deviations from the long-term

equilibrium relationship, allowing for the analysis of dynamic adjustments when variables temporarily move away from their long-term paths.

4. **Granger Causality Testing:**

   Cointegration is essential when conducting Granger causality tests between variables. In the presence of a cointegrated relationship, it becomes possible to distinguish between short-term and long-term causality, providing more accurate insights into the nature of the relationships.

5. **Forecasting:**

   Cointegration improves the reliability of forecasts by considering the long-term relationships between variables. Once cointegrated relationships are identified, forecasting models can be built that take into account the stable, long-term behavior of the variables.

6. **Risk Management in Financial Markets:**

   In finance, cointegration is used to identify pairs of assets that move together in the long run, forming the basis for pairs trading strategies. Traders use cointegration to manage risk by anticipating mean-reverting behavior in asset prices.

7. **Policy Analysis:**

   Cointegration is employed in economic and policy analysis to study the relationships between key macroeconomic variables. For example, it helps economists understand the long-term equilibrium relationship between inflation and unemployment.

8. **Improved Statistical Inference:**

   Cointegration enhances the accuracy of statistical inference by ensuring that relationships identified between variables are not driven by spurious correlations due to non-stationarity.

By achieving these objectives, cointegration provides a robust framework

for analyzing and modeling relationships between economic, financial, and other time series variables, contributing to more accurate and meaningful conclusions in empirical studies.

19.4 IMPORTANCE OF COINTEGRATION

Cointegration is of significant importance in time series analysis, econometrics, and various fields due to several key reasons:

1. Handling Non-stationarity:

Many economic and financial time series data are non-stationary, meaning they exhibit trends or other patterns over time. Cointegration addresses this issue by identifying long-term relationships between non-stationary variables, allowing for more reliable analysis and modeling.

2. Long-Term Relationships:

Cointegration focuses on capturing long-term relationships between variables. This is crucial in understanding the fundamental links between economic or financial variables that persist over time, providing insights into the underlying economic mechanisms.

3. Error Correction Modeling:

Cointegration often involves the use of error correction models (ECM). These models account for short-term deviations from the long-term equilibrium, offering a dynamic framework for analyzing how variables adjust back to their long-term relationships after temporary shocks.

4. Granger Causality Testing:

Cointegration is essential for conducting Granger causality tests. It helps distinguish between short-term and long-term causality, improving the accuracy of causal relationships identified between variables.

5. Improved Forecasting:

Cointegration enhances forecasting accuracy by incorporating the stable, long-term relationships between variables. Forecasting models based

on cointegrated relationships are more reliable, especially in economic and financial contexts.

6. Risk Management in Financial Markets:

In finance, cointegration is valuable for identifying pairs of assets that move together in the long run. This forms the basis for pairs trading strategies, allowing investors and traders to manage risk by exploiting mean-reverting behavior in asset prices.

7. Portfolio Diversification:

Cointegration can be used to identify assets with stable long-term relationships, helping investors construct diversified portfolios. Understanding cointegrated relationships contributes to effective asset allocation strategies.

8. Policy Analysis:

Cointegration is employed in economic and policy analysis to study the relationships between key macroeconomic variables. For instance, it helps economists analyze the long-term equilibrium relationship between inflation and unemployment, providing insights for policy decisions.

9. Statistical Inference:

Cointegration improves the reliability of statistical inference. By accounting for the non-stationarity of variables, it helps ensure that identified relationships are not spurious and are statistically valid.

10. Pairs Trading Strategies:

In financial markets, cointegration is often used to identify pairs of assets that exhibit mean-reverting behavior. Traders can take advantage of temporary deviations from the long-term equilibrium by going long on the undervalued asset and short on the overvalued asset.

In summary, cointegration is crucial for addressing the challenges posed by non-stationarity in time series data, providing a framework for understanding

and modeling long-term relationships, improving forecasting accuracy, and supporting various applications in finance, economics and beyond.

19.5 DESCRIBE THE DETERMINISTIC TRENDS

Deterministic trends refer to systematic and predictable patterns or movements in time series data that follow a specific mathematical function or form. Unlike stochastic or random trends, which exhibit unpredictable fluctuations, deterministic trends show a clear and consistent direction over time. These trends are often modeled using mathematical equations to capture the underlying structure of the data.

There are different types of deterministic trends, each characterized by its mathematical representation. Here are three common forms:

1. Linear Trend:

A linear trend is a straight-line movement in the data over time.

The mathematical form of a linear trend is given by the equation: $Yt = \alpha + \beta t + \varepsilon t$, where Yt is the value of the time series at time $t\alpha$ is the intercept, β is the slope or rate of change, t is time, and εt is the error term.

2. Quadratic Trend:

A quadratic trend represents a curved movement in the data over time.

The mathematical form of a quadratic trend is given by the equation: $Yt = t\alpha + \beta 1 + \beta 2t2\varepsilon t$, where Yt is the value of the time series at time t, α is the intercept, $1\beta 1$ and $2\beta 2$ are coefficients, t is time, and εt is the error term.

3. Exponential Trend:

An exponential trend reflects a consistent rate of growth or decay over time.

The mathematical form of an exponential trend is given by the equation: $Yt = \alpha + e\beta t + \varepsilon t$, where Yt is the value of the time series at time t, α is

a scaling factor, β is the growth or decay rate, t is time, and εt is the error term.

Characteristics of Deterministic Trends:

1. **Predictability:** Deterministic trends are predictable and follow a specific mathematical pattern, making it easier to forecast future values based on historical data.
2. **Systematic Movement:** The movements in deterministic trends are systematic and driven by a consistent mathematical relationship. There is no randomness or unpredictability associated with these trends.
3. **Noisy Components:** Deterministic trends may still have random or noisy components represented by the error term (εt), which captures unobservable factors and measurement errors.
4. **Stationarity:** Deterministic trends are often associated with non-stationary time series, as they do not exhibit constant statistical properties over time.

These deterministic trends are important in time series analysis as they provide a structured way to understand and model the underlying patterns in the data. Identifying and accounting for deterministic trends is crucial for accurate forecasting, trend analysis, and decision-making in various fields such as economics, finance, and environmental science.

19.6 IMPORTANCE OF DETERMINISTIC TRENDS

Deterministic trends are important in time series analysis for several reasons, spanning various fields such as economics, finance, environmental science, and more. Here are some key aspects highlighting the importance of deterministic trends:

1. **Forecasting and Prediction:**

   Deterministic trends provide a structured framework for forecasting future values of a time series. By identifying and modeling these trends,

analysts can make predictions about the future direction and magnitude of the variable, aiding in decision-making and planning.

2. **Policy and Strategy Formulation:**

In economics and finance, understanding deterministic trends is crucial for policymakers, businesses, and investors. Decision-makers can use trend analysis to formulate effective policies, investment strategies, and business plans based on the expected direction of key economic indicators.

3. **Long-Term Planning:**

Deterministic trends help in long-term planning by providing insights into the persistent patterns or movements in a variable. This is valuable for industries, governments, and organizations that need to make strategic decisions with a long-term perspective.

4. **Identifying Structural Changes:**

Changes in deterministic trends can signal structural shifts in the underlying dynamics of a system. Detecting such changes is essential for understanding evolving patterns and adjusting strategies accordingly.

5. **Economic and Financial Modeling:**

Deterministic trends are fundamental components in economic and financial models. They are used to represent the systematic, non-random movements in economic indicators, prices, and other financial variables.

6. **Environmental and Climate Analysis:**

In environmental science, deterministic trends play a role in analyzing long-term changes in climate patterns, temperature, and other environmental variables. This information is crucial for understanding the impact of climate change and developing mitigation or adaptation strategies.

7. **Resource Allocation:**

Businesses and governments often allocate resources based on expected

trends. Deterministic trend analysis helps in optimizing resource allocation by providing a clearer picture of the expected future values of relevant variables.

8. Risk Management:

Deterministic trends are valuable in risk management, especially in finance. Understanding the long-term movements of asset prices helps investors and financial institutions make informed decisions about risk exposure, portfolio diversification, and hedging strategies.

9. Time Series Decomposition:

Decomposing a time series into its deterministic trend component, along with other components like seasonality and residuals, aids in understanding the underlying structure of the data. This decomposition is valuable for advanced time series analysis and modeling.

10. Infrastructure Planning:

For sectors like transportation, energy, and urban planning, deterministic trends are essential for designing and developing infrastructure projects that meet long-term demands and accommodate growth.

In summary, deterministic trends provide a foundational understanding of the systematic movements in time series data, enabling better decision-making, forecasting, and planning across a range of disciplines. Their importance lies in their ability to capture persistent patterns and contribute to more accurate and insightful analyses of dynamic systems.

19.7 DESCRIBE THE STOCHASTIC TRENDS

Stochastic trends, in contrast to deterministic trends, represent random and unpredictable fluctuations in time series data. These trends are characterized by movements that do not follow a specific mathematical pattern or deterministic function. Stochastic trends introduce an element of randomness into the data, making them essential to understanding the inherent uncertainty and variability in various phenomena. Here are key features and aspects related to stochastic trends:

1. **Random and Unpredictable:**

   Stochastic trends exhibit random movements that cannot be precisely predicted based on historical data. Unlike deterministic trends, there is no systematic mathematical relationship governing the direction or magnitude of these fluctuations.

2. **Stationary and Non-Stationary Components:**

   Stochastic trends are often associated with non-stationary time series data. Non-stationary series have statistical properties that change over time, and stochastic trends contribute to this non-stationarity. However, the presence of a stochastic trend does not imply non-stationarity by itself.

3. **Unit Root:**

   Stochastic trends are often associated with the concept of a unit root in time series analysis. A unit root indicates that a time series has a stochastic trend and lacks a stable, long-term mean.

4. **Long-Term Persistence:**

   Stochastic trends may exhibit long-term persistence, meaning that the effects of a shock to the system can last for an extended period. This characteristic makes it challenging to distinguish between temporary fluctuations and more persistent changes in the data.

5. **Random Walk Models:**

   A common representation of stochastic trends is found in random walk models. In a random walk, each observation is the result of a random shock or innovation, and future values are unpredictable based on past observations. Brownian motion is a continuous-time stochastic process often used to model random walks.

6. **Cointegration with Deterministic Trends:**

   In some cases, a combination of stochastic and deterministic trends may exist in a time series. Cointegration is a concept used to model

relationships between non-stationary variables, where a linear combination of the variables forms a stationary series.

7. Economic and Financial Applications:

Stochastic trends play a significant role in economic and financial modeling. They capture the inherent uncertainty and random shocks that influence economic variables, asset prices, and other financial indicators.

8. Policy Uncertainty:

The presence of stochastic trends highlights the uncertainty inherent in economic and policy-related variables. Policymakers need to consider the random nature of shocks when formulating strategies and making decisions.

9. Monte Carlo Simulations:

Stochastic trends are commonly used in Monte Carlo simulations, a method that involves generating a large number of random scenarios to analyze the possible outcomes of a system. This is valuable in assessing risk and uncertainty in various fields.

Understanding stochastic trends is crucial for accurate modeling and forecasting in situations where randomness and unpredictability play a significant role. It is a key concept in time series analysis, econometrics, and various scientific disciplines where capturing the inherent variability in data is essential for making informed decisions.

19.8 IMPORTANCE OF THE STOCHASTIC TRENDS

Stochastic trends are important in various fields, including economics, finance, environmental science, and more. Their significance lies in their ability to capture random fluctuations and uncertainties in time series data. Here are several reasons highlighting the importance of stochastic trends:

1. Realism in Economic and Financial Models:

Stochastic trends add a realistic element to economic and financial

models by acknowledging the inherent uncertainty and randomness in economic variables, asset prices, and other financial indicators. This realism is crucial for making models that better reflect the complexities of the real world.

2. Risk Management in Finance:

In finance, understanding stochastic trends is essential for risk management. Stochastic fluctuations represent the random shocks and uncertainties that financial instruments face, helping investors and financial institutions assess and manage risks associated with market volatility.

3. Option Pricing Models:

Stochastic processes, such as geometric Brownian motion, are used in option pricing models. These models take into account the random movements of asset prices, providing a more accurate representation of the uncertainty involved in option pricing.

4. Monte Carlo Simulations:

Stochastic trends play a crucial role in Monte Carlo simulations, where random scenarios are generated to analyze the potential outcomes of a system. This method is widely used in various fields, including finance, engineering, and operations research, for risk assessment and decision-making.

5. Understanding Economic Fluctuations:

Stochastic trends contribute to understanding the fluctuations in economic indicators that cannot be explained solely by deterministic trends. This understanding is vital for policymakers, economists, and businesses in responding to economic uncertainties and planning for various scenarios.

6. Predicting Macroeconomic Variables:

Stochastic trends help model and predict macroeconomic variables, such

as GDP, inflation, and unemployment, by accounting for the random shocks that these variables may experience. This is important for making informed policy decisions and assessing the overall health of an economy.

7. **Energy and Environmental Modeling:**

Stochastic trends are employed in modeling energy prices, environmental variables, and climate patterns. These models consider the random fluctuations that impact energy markets and environmental systems, aiding in long-term planning and decision-making.

8. **Improved Time Series Analysis:**

Incorporating stochastic trends in time series analysis allows researchers and analysts to better capture the complexity of real-world data. This leads to more accurate statistical models and predictions, especially in cases where deterministic trends alone may not suffice.

9. **Investment Strategies:**

Investors use stochastic trends to develop investment strategies that account for the uncertainty and randomness in financial markets. Strategies that take into consideration stochastic fluctuations can be more robust and adaptable to changing market conditions.

10. **Policy Formulation:**

Policymakers consider stochastic trends when formulating economic and fiscal policies. Understanding the random nature of economic shocks helps policymakers design more resilient policies that can withstand unforeseen events.

In summary, stochastic trends are important for providing a more realistic representation of the uncertainties and random fluctuations present in various systems. Incorporating these trends into models and analyses leads to more accurate predictions, better risk management, and improved decision-making in a wide range of fields.

19.9 DESCRIBE THE UNIT ROOT TESTS

Unit root tests are statistical tools used in time series analysis to determine whether a time series variable is non-stationary and exhibits a unit root. A unit root implies that the time series has a stochastic or random trend and lacks a stable, long-term mean. Unit root tests are crucial in understanding the behavior of time series data, especially in econometrics and finance. Here are some key aspects of unit root tests:

1. Concept of a Unit Root:

A unit root is a characteristic of a non-stationary time series. If a time series has a unit root, it means that the series does not revert to a stable mean over time. Instead, it exhibits a random walk or stochastic trend, making it non-stationary.

2. Null Hypothesis and Alternative Hypothesis:

The null hypothesis ($0H0$) in unit root tests is that a unit root is present, indicating non-stationarity. The alternative hypothesis ($1H1$) is that the unit root is absent, suggesting stationarity.

3. Common Unit Root Tests:

3.1 Augmented Dickey-Fuller (ADF) Test:

The ADF test is one of the most widely used unit root tests. It extends the Dickey-Fuller test by including lagged differences of the time series variable. The test statistic is compared to critical values to determine whether to reject the null hypothesis.

3.2 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test:

The KPSS test is an alternative to the ADF test. It focuses on the null hypothesis of stationarity, and rejection of the null implies the presence of a unit root.

3.3 Phillips-Perron (PP) Test:

Similar to the ADF test, the PP test is used to check for the presence of a unit root. It allows for certain types of serial correlation in the errors.

4. **Critical Values and Decision Rule:**

Unit root tests compare the calculated test statistic to critical values from statistical tables or obtained through simulations. If the test statistic exceeds the critical value, the null hypothesis of a unit root is rejected, indicating stationarity.

5. **Dickey-Fuller Test (DF):**

The Dickey-Fuller test is a special case of the ADF test when lag order is set to zero. It is less powerful than the ADF test but is often used for simplicity.

6. **Seasonal Unit Root Tests:**

Some unit root tests are designed to detect seasonal unit roots, which occur when a time series exhibits a unit root at a seasonal frequency. The Canova-Hansen and HEGY tests are examples.

7. **Testing for Structural Breaks:**

Unit root tests may be extended to account for structural breaks in the time series. Structural breaks can affect the stationarity properties of the data.

8. **Applications in Cointegration:**

Unit root tests are closely related to cointegration analysis. If two or more non-stationary time series are found to be cointegrated, it implies the existence of a long-term relationship, and unit root tests help confirm the non-stationarity of the individual series.

Unit root tests are valuable tools for detecting non-stationarity in time series data, helping researchers and analysts choose appropriate modeling techniques and improve the accuracy of their analyses, especially in forecasting and econometric applications.

19.10 SUMMARY OF RANDOM WALK MODELS

Definition: Random walk models describe the stochastic or random

movement of a variable over time. The future values of the variable are determined by a series of random steps, and each step is independent of previous ones.

Types of Random Walk Models:

1. **Discrete Random Walk:**
   - Variable changes at discrete time intervals.
   - Commonly used in one-dimensional scenarios, where the variable can move in one of several directions with equal probability.
2. **Continuous Random Walk:**
   - Variable changes continuously over time.
   - Modeled using stochastic differential equations.
   - Brownian motion is a specific continuous random walk widely used in physics and finance.

Key Characteristics:

1. **Markov Property:**
   - Future values depend only on the current state and are independent of past states.
2. **No Memory:**
   - Each step is independent of previous steps, exhibiting no memory of past movements.
3. **Diffusion-Like Behavior:**
   - Random walks often result in a diffusion-like behavior, where values spread out over time.

Applications:

1. **Financial Markets:**
   - Used to model stock prices, assuming that price changes are unpredictable and follow a random pattern.

2. Physics:

- Brownian motion models the random movement of particles in a fluid, with applications in statistical physics.

3. Biology:

- Applied to describe biological processes like the movement of cells or organisms in response to random stimuli.

Limitations:

1. Simplification:

- Random walk models are simplifications that may not capture all aspects of complex systems.

2. Efficient Market Hypothesis (EMH):

- Assumes that stock prices follow a random walk, contributing to the development of the Efficient Market Hypothesis.

Conclusion: Random walk models provide a valuable framework for understanding the stochastic behavior of variables over time. While they offer simplicity and insights into certain phenomena, their limitations and assumptions need to be considered, especially in contexts where factors beyond randomness significantly influence the dynamics of the system.

19.11 GLOSSARY

- **Cointegration:** Cointegration is a statistical property that describes a long-term relationship between two or more non-stationary time series variables. When two or more series are cointegrated, it means that although each series may individually follow a random walk or be non-stationary, there exists a linear combination of them that is stationary. In simpler terms, cointegration implies that the variables share a common stochastic trend.
- **Deterministic trends:** A deterministic trend refers to a systematic and predictable long-term pattern or movement in a time series that follows

a clear and consistent trajectory. Unlike stochastic (random) trends, which are unpredictable and may vary over time, deterministic trends exhibit a constant and regular behavior.

- **Unit root tests:**Unit root tests are statistical tests used in time series analysis to assess whether a time series variable has a unit root or not. A unit root is a characteristic of a non-stationary time series where the process has a stochastic (random) trend, meaning it doesn't revert to a constant mean over time. In other words, a unit root implies that the time series variable follows a random walk or a non-stationary process.Unit root tests are essential in determining the stationarity of a time series because many traditional time series models assume or work best with stationary data. A unit root indicates that differencing (a common technique to induce stationarity) is required for the time series to be stationary.

19.12 SELF ASSESSMENT QUESTIONS

1. Explain the concept of Unit root tests?

_____

_____

_____

2. Explain how the stochastic trends are important in Econometrics?

_____

_____

_____

3. Explain the three Objectives of cointegration?

_____

_____

_____

19.13 SUGGESTED READINGS

1. Aileen Nielsen, 2019"Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics",4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell*. *Andrew W*. *Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*
10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"
11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"
12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"
13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Ruud, Paul A., 2000 “An Introduction to Classical Econometric Theory”, Oxford, 2000.

15. Wooldridge, J M, 2009 “Introductory Econometrics – A Modern Approach” (4th ed), South-Western, 2009.

******

UNIT-IV **LESSON NO. 20**

TIME SERIES ECONOMETRICS

APPROACHES TO ECONOMETRIC FORECASTING

STRUCTURE

20.1 INTRODUCTION

Forecasting is simply the process of using past data values to make educated predictions on future data values. As stated in the last chapter, the time series should be stationary if you want to make well-informed predictions. This can be done by fitting an arima model by using the *auto.plot ()* function in the **forecast** package. Then, all you have to do is apply the *forecast ()* function

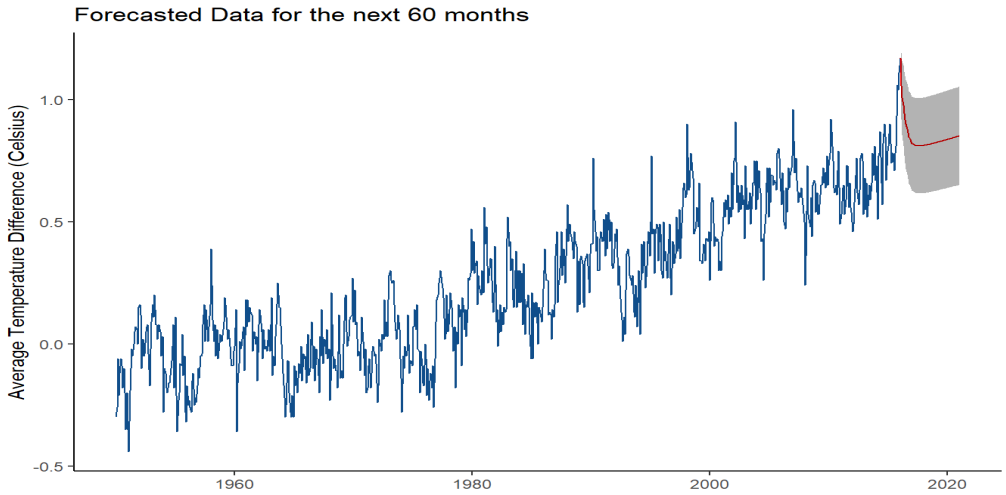
to get your prediction! The *forecast ()* function can take in another argument along with your model. You can also input h, the number of predicted time periods you want. This function is very practical for real world analysis of time series.

Forecasting is done in so many fields around the world. You will often see forecasting in the business and financial field for companies that want to predict their profit or expenses. Forecasting can be used to predict stock prices as well! You will see it in the environmental field, such as this current example with global warming. The economic field also heavily uses time series and forecasting to predict how societies will behave. This is just a few examples of numerous time series and forecasting uses in the real world.

Example: Global Temperature

Let's forecast with our global temperature data now. As we saw, we fit the data with a SARIMA (2,1,3)(1,0,0)12. Now that we have our model, we can simply use the forecast (ts,h) function from the **forecast** package. As mentioned, h represents the number of observations we want to predict into the future. Let's say we want to predict 5 years into the future. Since our data ended in December 2016, the next 5 years will include each month from January 2017 to December 2022. Moreover, since our observations were every month, we can set h to 12"5=6012"5=60. Let's see what happens.

```
forecast.data <-forecast(best.model, h =60)
autoplot(forecast.data, ts.colour =”dodgerblue4", predict.colour=”red”) +
ggtitle (“Forecasted Data for the next 60 months”) +
ylab(“Average Temperature Difference (Celsius)”) +
theme_classic()
```



The original time series is depicted in dark blue and the predicted data is represented by the red line. The grey shading around the predicted values represents the 95% confidence interval. This simply states that we are 95% percent confident that the data point at time *t* will fall between two bounds. You can change the confidence bands in the forecast function and even view the confidence interval bounds in the forecasted object.

What can you take away from this forecasted model?

It looks as it the average temperature difference will drop in 2017; however, it seems that it will gradually rise again. This is fairly consistent in what we see in the rest of the data. It seems in the past 66 years, the temperature decreases every once in a while, but then gradually rises. This looks like it's a fairly good prediction.

20.2 APPROACHES TO ECONOMETRIC FORECASTING

Econometric forecasting involves using statistical and mathematical models to make predictions about future economic conditions. Several approaches are commonly used in econometrics for forecasting. Here are some of the key approaches:

1. **Time Series Analysis:**

   Description: This approach focuses on analyzing historical time-ordered data to identify patterns, trends, and seasonality. Time series models, such as ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), and exponential smoothing, are commonly used for forecasting economic variables over time.

2. **Regression Analysis:**

   Description: Regression models establish relationships between dependent and independent variables. In econometric forecasting, regression analysis is often used to model the impact of various factors on economic variables. Multiple regression models can incorporate several predictors to enhance forecasting accuracy.

3. **VAR (Vector Autoregression) Models:**

   Description: VAR models are designed to capture the dynamic relationships among multiple time series variables. These models are particularly useful for forecasting when variables are interrelated and influence each other. Granger causality tests and impulse response functions are common tools in VAR analysis.

4. **Cointegration Analysis:**

   Description: Cointegration explores long-term relationships between non-stationary time series variables. It is often used in combination with vector error correction models (VECM) to analyze and forecast variables that are bound together in the long run.

5. **Bayesian Econometrics:**

   Description: Bayesian econometrics incorporates Bayesian principles into econometric models. Bayesian methods allow for the incorporation of prior information and updating beliefs based on new data. Bayesian econometrics is useful when dealing with small sample sizes or when there is significant uncertainty about parameter values.

6. **Machine Learning Models:**

   Description: Machine learning techniques, such as neural networks, support vector machines, and random forests, are increasingly used in econometric forecasting. These models can capture complex patterns and relationships in data, especially when dealing with large datasets and nonlinear relationships.

7. **Agent-Based Modeling:**

   Description: Agent-based models simulate the interactions of individual agents within an economic system. This approach is useful for capturing emergent properties and complex interactions that may not be easily modeled using traditional econometric techniques.

8. **Forecast Combination:**

   Description: Forecast combination involves aggregating forecasts from multiple models to improve overall accuracy. This approach recognizes that different models may perform well under different conditions, and combining their predictions can lead to more robust forecasts.

9. **Ensemble Methods:**

   Description: Ensemble methods, such as bagging and boosting, combine multiple models to create a more accurate and stable forecast. By leveraging the strengths of different models, ensemble methods aim to produce a more reliable and robust forecasting outcome.

Considerations:

Data Quality: The quality and reliability of the data used for forecasting are critical.

Model Validation: Forecasts should be validated using out-of-sample data to assess their accuracy and generalizability.

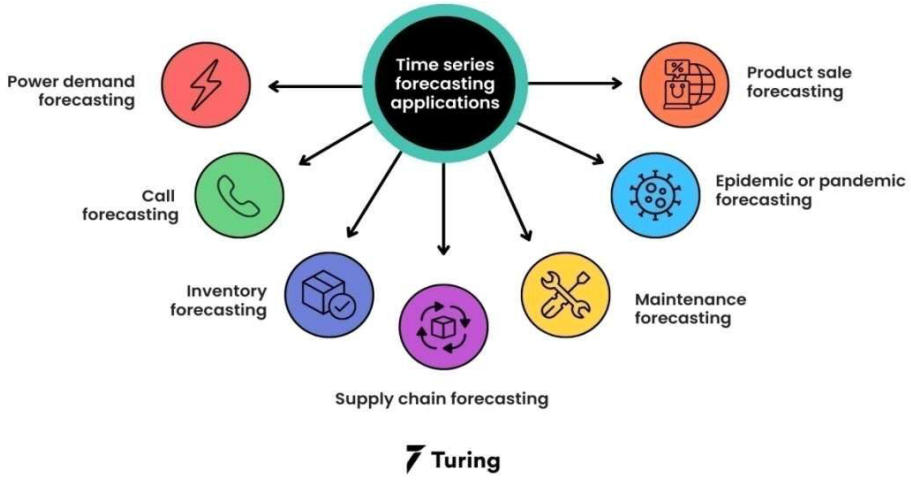
Assumptions: Assumptions underlying the chosen econometric models should be carefully considered and tested.

The choice of the forecasting approach depends on the nature of the data, the specific economic variables of interest, and the characteristics of the problem at hand. Often, a combination of approaches or model averaging techniques is employed for more robust forecasting results.

20.3 WHAT IS TIME SERIES FORECASTING ?

In time series forecasting, time series data is analyzed through statistics and mathematical modeling to predict and inform strategic decisions. It's neither an exact prediction nor is it possible to predict with 100% accuracy, particularly while dealing with frequently changing variables and some beyond our control variables. However, forecasting can provide insight into the likelihood of certain outcomes. Generally, a more extensive dataset leads to more accurate forecasting. Predictions and forecasts are generally synonymous, but there is a notable difference between them. Prediction refers to data at a general future point in time, whereas forecasting focuses on data at a specific future point when it occurs. The analysis of time series is often combined with series forecasting. In time series analysis, models are developed to understand the underlying causes of the data. By analyzing outcomes, you can understand "why" they occur. As a result, forecasting takes the next step of extrapolating the future from the knowledge derived from the past.

Applications of time series forecasting



From sales forecasting to weather forecasting, time series models have a wide range of applications.

A time series model is one of the most effective methods of forecasting when there is uncertainty about the future.

A time series forecast plays a crucial role in every category of the business decision. Here are some examples:

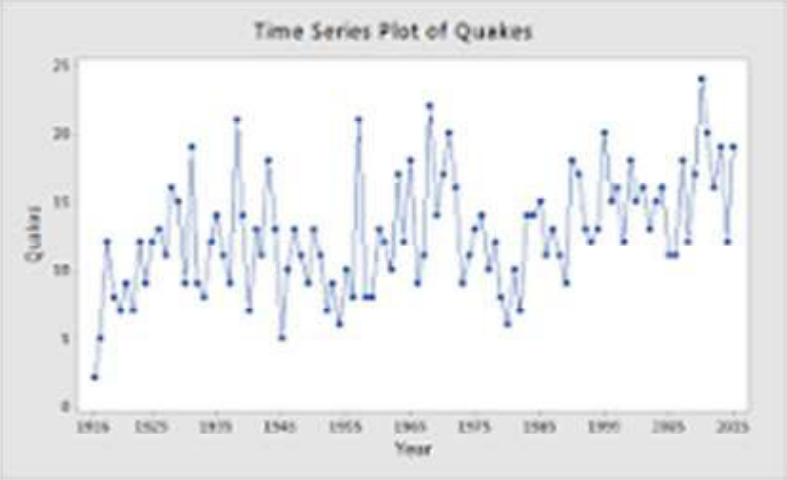
- Power demand forecasts help determine whether another power plant should get built in the next five years.
- Scheduling of calls for the next week is done based on the call volume forecast.
- To forecast inventory requirements so as to stock items to meet demand.
- A supply chain management forecast is used to optimize fleet management and other aspects of the supply chain.
- To minimize downtime and maintain safety standards by predicting equipment failures and maintenance requirements.
- The outbreak of epidemic or pandemic is controlled by forecasting infection rates.
- Analyzing customer ratings and forecasting product sales.

Different forecasts involve different time horizons and can depend on the circumstances

20.4 WHAT IS THE AR MODEL OF TIME SERIES?

A time series is a sequence of measurements of the same variable(s) made over time. An autoregressive model is when a value from a time series is regressed on previous values from that same time series.

For example, on y $yt-1: yt=\beta 0 yt-1+\varepsilon t$



20.5 AR MODEL VS MA MODEL

In time series analysis, the Autoregressive (AR) model and the Moving Average (MA) model are foundational concepts. They are often combined to form more sophisticated models like ARMA and ARIMA. Let's delve into each.

1. Autoregressive (AR) Model:

An autoregressive model predicts a value in a time series using a linear combination of past values of the series. The term "autoregressive" indicates that it's a regression of the variable against itself.

The AR(p) model is defined as:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t$$

Where:

- X_t is the value of the series at time t.
- c is a constant.
- p is the order of the AR model, indicating how many lagged past values are used.
- $\phi_1, \phi_2, \ldots$ are the parameters of the model.
- ϵ_t is white noise or error.

2. Moving Average (MA) Model

The moving average model, in the context of time series (not to be confused with the simple moving average), uses past white noise terms (or error terms) to predict the series.

The MA(q) model is defined as:

$$X_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q}$$

Where:

- X_t is the value of the series at time t.
- μ is the mean of the series.
- q is the order of the MA model, indicating how many lagged error terms are used.
- $\theta_1, \theta_2, ...$ are the parameters of the model.
- ϵ_t is white noise or error.

20.6 HOW THEY DIFFER

- The AR model relates the current value of the series to its past values. It assumes that past values have a linear relationship with the current value.
- The MA model relates the current value of the series to past white noise or error terms. It captures the shocks or unexpected events in the past that are still affecting the series.

Combined Models:

Often, these models are combined to model and forecast time series data more effectively:

- ARMA (Autoregressive Moving Average): This model combines both AR and MA components.

- ARIMA (Autoregressive Integrated Moving Average): This model adds an "I" (integrated) component, which involves differencing the series to make it stationary before applying an ARMA model.

Both AR and MA models (and their combinations) are foundational in time series forecasting, and their applicability depends on the characteristics of the data and the nature of the underlying processes generating the time series.

20.7 SUMMARY

A time series is a series of data points over time. Trend, seasonality, random white noise. Stationarity is defined as constant means and variance throughout the series. Time series should be stationary when forecasting. Must fit a model to the time series to detrend and deseasonalize the series. The autoregressive model uses observations from previous time steps as input to regression equations to predict the value at the next step. The moving average model is a time series model that accounts for very short-run autocorrelation. You should use the ACF and PACF plots to determine the order of these. These models only work on stationary data. ARMA models are a combination of AR and MA models and only works on stationary data. ARIMA models are the same as an ARMA model, but there is a differencing term to detrend the non-stationary data. SARIMA models are the same as an ARIMA model, but there are seasonal terms to detrend and deseasonalize the non-stationary data. Forecasting is defined as using the previous data points to make well-informed predictions of the future, The forecast () function is best to predict your time series. These is just a small chunk of what time series has to offer. There are many other components and different techniques to detrend and deseasonalize your time series that we didn't talk about in this book. I hope this short tutorial assisted you in learning the basics of time series and equipped you with the tools to forecast. Now that you have learned the basics, make sure to check out our corresponding API package and Stock Shiny App to explore real world uses of time series. You will be able to look at the series of real time stock prices for companies around the world!

ARMA, ARIMA, and SARIMA Models

ARMA models are a combination of AR and MA models and only works on stationary data.

ARIMA models are the same as an ARMA model, but there is a differencing term to detrend the non-stationary data.

SARIMA models are the same as an ARIMA model, but there are seasonal terms to detrend and deseasonalize the non-stationary data.

In practice, use the auto.arima () function from the forecast package.

Forecasting

Defined as using the previous data points to make a well-informed predictions of the future.

The **forecast ()** function is best to predict your time series.

These is just a small chunk of what time series has to offer. There are many other components and different techniques to detrend and deseasonalize your time series that we didn't talk about in this book. I hope this short tutorial assisted you in learning the basics of time series and equipped you with the tools to forecast.

Now that you have learned the basics, make sure to check out our corresponding API package and Stock Shiny App to explore real world uses of time series You will be able to look at the series of real time stock prices for companies around the world!

20.8 GLOSSARY

An autoregressive (AR) model is a type of time series model that describes the dependence between an observation and its previous observations. In an AR model, the current value of the time series is modeled as a linear combination of its past values, plus a random error term. The "autoregressive" nature of the model comes from the fact that the regression is performed on the series itself, creating a relationship between an observation and its lagged (past) value. Time series forecasting refers to the process of predicting future values

of a variable based on its historical values and patterns. In other words, it involves using past observations to make informed predictions about the future behavior of a time-ordered sequence of data points. Time series forecasting is commonly applied in various fields, including finance, economics, meteorology, engineering, and many others. Approaches to econometric forecasting refer to the various methods and techniques used in econometrics to predict future economic variables based on historical data and statistical models. Econometric forecasting involves the application of statistical and mathematical models to analyze economic relationships, estimate parameters, and make predictions about future economic outcomes. There are several approaches to econometric forecasting, each with its strengths, assumptions, and suitability for different types of data and economic phenomena. Here are some common approaches: (i.) Time Series Models. (ii.) Causal Models. (iii.) Leading Indicators. (iv.) Machine Learning Models. (v.) Bayesian Econometrics. (vi.) Forecast Combination. (vii.) Judgmental Forecasting.

20.9 SELF ASSESSMENT QUESTIONS

1. Explain the approaches to econometric Forecasting?

_____

_____

_____

2. Explain the concept of time series forecasting?

_____

_____

_____

3. Explain the difference between AR model and MA model?

_____

_____

_____

20.10 SUGGESTED READINGS

1. Aileen Nielsen, 2019 "Practical Time Series Analysis" Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041658.
2. Aronow, P. M., and B. T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press. https://books.google.co.uk/books?id=u1N-DwAAQBAJ
3. Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, "Introduction to Time Series Analysis and Forecasting"
4. Greene, William H., Econometric Analysis, Prentice Hall, 2000.
5. Econometrics by Damodar Gujarati
6. Econometric Analysis, Willam H. Greene, Stern School of Business, New York University
7. Hill R.C., Griffiths W.E., Lim G.C. "Principles of Econometrics",4th edition, [ISBN 978-1-11803207-*7*
8. Jeffrey M. Wooldridge "Introductory Econometrics A Modern Approach", *6th Edition* - 8 October 2015. ISBN-13: 978-1305270107
9. *John Y. Campbell*. *Andrew W. Lo*. A. *Craig MacKinlay,* 1996 "The Econometrics of Financial Markets" ISBN: 9780691043012; Published: *Dec 29*, *1996*
10. Robert H. Shumway, David S. Stoffer" Characteristics of Time Series"
11. Robert H. Shumway, David S. Stoffer "Time Series Regression and Exploratory Data Analysis"
12. Robert H. Shumway, David S. Stoffer "Statistical Methods in the Frequency Domain"
13. Robert H. Shumway , David S. Stoffer "Time Series Analysis and Its ApplicationsWith R Examples"

14. Ruud, Paul A., 2000 “An Introduction to Classical Econometric Theory”, Oxford, 2000.

15. Wooldridge, J M, 2009 “Introductory Econometrics – A Modern Approach” (4th ed), South-Western, 2009.

****