Unit: 1-9

Course Code: 9313

Credit Hours: 3

AIOU

FUNDAMENTAL OF ECONOMETRICS

B.S. Economics (4 Years)

ALLAMA IQBAL OPEN UNIVERSITY

www.aiou.edu.pk

STUDY GUIDE

FUNDAMENTALS OF ECONOMETRICS

BS ECONOMICS

Course Code: ECO 6003/ 9313 **Units: 1–9**

Credit Hours: 03

Department of Economics
Faculty of Social Sciences & Humanities
Allama Iqbal Open University, Islamabad

1st Edition…….2023

Quantity……1000

Layout :……..Naeem Akhtar

Printing Coordinator:……Dr. Sarmad Iqbal

Printer……AIOU, Islamabad

Publisher……. AIOU, Islamabad

COURSE TEAM

Course Team:	Dr. Fouzia Jamshaid Dr. Muhammad Ilyas Mr. Rizwan Ahmed Satti
Course Development Coordinator:	Mr. Rizwan Ahmed Satti
Writer:	Dr. Muhammad Jamil
Reviewer:	Mr. Rizwan Ahmed Satti
Editor:	Mr. Fazal Karim
Composing & Layout:	Naeem Akhtar

CONTENTS

PREFACE

The curriculum at AIOU is designed on modern parameters using the latest information, trends, theories, and techniques. An extensive consultative process is also a basic component of the activity. Development of the study material to help the students located throughout the country is taken as a challenge. AIOU takes pride in undertaking this major task for the effective learning of the students.

The BS Economics is being offered by the Department of Economics of Allama Iqbal University for the students who are interested in the field of economics. The scheme of study for BS Economics has been designed and the courses are developed to make these relevant to the emerging national and global trends and to meet needs of the society in this domain. The study material provides a comprehensive coverage of the core contents for BS Economics Program. The selection of, and the treatment of study materials have been designed to meet both the general and specific aims set out by the Higher Education Commission (HEC) through the National Curriculum Revision Committee (NCRC).

In the end, I am happy to extend my gratitude to the Course Team, Chairman, Course Development Coordinator, Unit-writer, reviewer, and Editor for the development of the course. Any suggestions for the improvement in the course will be warmly welcomed by the Department of Economics.

Prof. Dr. Nasir Mehmood
Vice Chancellor

INTRODUCTION TO THE COURSE

Welcome to course "Fundamentals of Econometrics" which is part of BS Economics scheme of study of the Department of Economics, Faculty of Social Sciences and Humanities, Allama Iqbal Open University Islamabad Pakistan.

Fundamentals of Econometrics is a basic course of econometrics which is concerned with the tolls of economic theory, mathematics, and statistical inference that are applied to the analysis of economic phenomenon. It consists of a set of tools to help in understanding the subject matter of econometrics and the concept of regression models. It also provides some tools of estimation and diagnostic tests for model selection. Fundamentals of Econometrics is one of the most important course in the subject of Economics which deals with the empirical analysis of economic models and forecasting.

This course will provide the concepts of econometrics, two variable regression models and multiple variables regression models. The subject matter handled in this course may be classified into broad sub-groups of Econometrics: specially concepts of regression analysis, Heteroskedasticity, Multicollinearity and Autocorrelation. There are nine units in total. The first unit is devoted to the concepts of Introduction to Econometrics. The rest of the units linked to each other. In each of broad areas tackled, the format adopted is as follow:

Unit two covers the analysis of two variables regression model, its estimation and statistical inference. It also covers the properties of the least squares method and the measurement of goodness of fit. Unit three presents the multiple regression model, its estimation, and tests for stability. Unit four consists of the matrix approach to linear regression model, hypothesis testing and the correlation Matrix. In unit five the concepts of multicollinearity will be discussed. Units six and seven consist of the concepts of Heteroskedasticity and Autocorrelation respectively. Unit eight presents a comprehensive account of Model specification and Diagnostic Testing. In unit nine Simultaneous Equation Models will be discussed.

The Study Guide in your hand provides you with the introduction to each Unit followed by the objectives of the Unit. In each Unit throughout the Study Guide, we have been given self-assessment questions. They are meant to assist your comprehension after reading the unit. A useful reading list is also provided for each Unit.

This is a Fundamental level 03 credit hours course on Econometrics, specially designed for students through the distance education system of the Allama Iqbal Open University. This course will be presented through the Learning Management System (LMS) Aaghi Portal. We hope that you will find this course useful and interesting. Suggestions for the improvement of course, as well as the Study Guide, will be highly appreciated.

Rizwan Ahmed Satti
Course Coordinator

COURSE LEARNING OUTCOMES

The course is designed for use in the semester study of Fundamentals of Econometrics. It should also be useful to everyone who seeks to address the question of relevance of Econometrics.

Upon successful completion of this course, the learner will be able to.

- Explain the concept of Econometrics.
- Describe the Two Variables and Multivariable Regression Analysis.
- Understand the concepts of Multicollinearity, Heteroskedasticity and Auto correlation.
- Analyze the structure of Model Specification and Diagnostic Testing.
- Comprehend the concept of Simultaneous Equation Models.

Throughout this course, you will also see related learning outcomes identified in each unit. You can use the learning outcomes to help organize your learning and gauge your progress.

STRUCTURE OF THE STUDY GUIDE

The course "Fundamentals of Econometrics" a three credit hours course consists of nine units. A unit is a study of 12–16 hours of course work for two weeks. The course work of one unit will include study of compulsory reading materials and suggested books. You should make a timetable for studies to complete the work within the allocated time.

This study guide/course has been organized to enable you to acquire the skill of self-learning. For each unit an introduction is given, to help you to develop an objective analysis of the major and sub-themes discussed in the prescribed reading materials. Besides this, learning outcomes of each unit are very specifically laid down to facilitate in developing logical analytical approach. Summary of main topics has also been included in the contents to understand the topics. We have given you self-assessments questions and activities which are not only meant to facilitate you in understanding the required reading materials but also to provide you an opportunity to assess yourself. Recommended books and important links have been given to understand the main topics. Key terms have also been included in the study guide.

Every course has a study package including study guides, assignments and tutorial schedule uploaded by the University. For the books suggested at the end of each unit, you can visit online resources, a nearby library, or the Central Library at the main campus in AIOU.

ABOUT THE AUTHOR(S)

All the units of the course “Fundamentals of Econometrics” are written by Dr. Muhammad Jamil who is serving as Professor of Economics, at Ghulam Ishaq Khan Memorial Chair (SBP), Kashmir Institute of Economics, The University of Azad Jammu & Kashmir, Muzaffarabad.

Dr. Muhammad Jamil's distinguished career in quantitative analysis and econometrics has been marked by his association with prestigious institutions and his profound expertise in the field. He had served as an Assistant Professor and later on the position of Associate Professor at the School of Economics, Quaid-i-Azam University, Islamabad, where he has taught various econometric courses and conducted workshops on applied economics. Dr. Jamil's teaching repertoire includes courses like "Econometric Methods," "Financial Econometrics," and "Time Series Econometrics," and he has been instrumental in training students and professionals in software such as EViews, RATS, Stata, SPSS, and LaTeX.

In addition to his teaching roles, Dr. Jamil's research and participation in seminars and workshops have made a significant impact on the field of econometrics in Pakistan. His research papers and active involvement in seminars on topics like branch-less banking reflect his ability to apply econometric techniques using various software tools. His work has not only contributed to the academic community, but has also played a vital role in advancing economic research and quantitative analysis in the institutions where he has served.

COURSE MATERIALS

The primary learning materials for this course are:

- Readings (e.g., study guides, recommended books, online links, and scholarly articles)
- Lectures, (workshops)
- Other resources.

All course materials are free to access and can be found through the links provided in each unit and sub-unit of the course. Pay close attention to the notes that accompany these course materials, as they will instruct you as to what specifically to read or watch at a given point in the course and help you to understand how these individual materials fit into the course. You can also access to a list of all the materials used in this course by clicking on resources mentioned in each unit.

Technical Requirements

This course is delivered online through the Learning Management System (LMS). You will be required to have access to a computer or web-capable mobile device and have consistent access to the internet either to view or download the necessary course resources and to attempt any auto-graded course assessments and the final exam.

Methods of Instruction

Following are the methods for directing this guide and course also and then you will be able to understand the macroeconomics course through.

- Lecture online
- Mandatory workshops
- Workshop Quizzes
- Class discussion during workshops
- Individual paired and small group exercises.
- Use of library for research projects
- Use of videos lectures
- Use of the internet

Types of Assignments

- Students must complete assignments from the recommended books and other sources also.
- Students must be able to research and complete the assignments, which will include library, Internet, and another media research.

Activities
In most of the units, different types of activates are mentioned for better understanding the course. If you thoroughly study the materials and follow the links and videos, then you will be able to understand the course in the easiest way.

HOW TO USE THE STUDY GUIDE?

Before attending a workshop, it is imperative to prepare yourself in the following manner to get the maximum benefit of it.
You are required to follow these steps:

Step 1
Go through the.
1. Course Outlines
2. Course Introduction
3. Course Learning Outcomes
4. Structure of the Course
5. Assessment Methods
6. Recommended Books
7. Suggested Readings

Step 2
Read the whole unit and make notes of those points which you could not fully understand or wish to discuss with your course tutor.

Step 3
Go through the self-assessment questions at the end of each unit. If you find any difficulty in comprehension or locating relevant material, discuss it online with your tutor.

Step 4
Study the compulsory recommended books at least for three hours in a week recommended in your study guide. AIOU Tries to read it with the help of a specific study guide for the course. You can raise questions on both during your online tutorial meetings and workshops.

Step 5
First go through assignments, which are mandatory to solve/complete for this course. Highlight all the points you consider difficult to tackle, and then discuss them in detail with your tutor. This exercise will keep you regular and ensure good results in the form of higher grades.

Assessment

For each three credit hours course, a student will be assessed as follow:

- Two Assignments (continuous assessment during semester).
- Final Examination (at the end of each semester)
- Mandatory participation in the workshop (as per AIOU policy)
- Workshop Quizzes
- Group discussion
- Presentation

Assignments

- Assignments are written exercises that are required to complete at home or place of work after having studied 9 units/study guides with the help of compulsory and suggested reading material within the scheduled study period. (See the assignments scheduled).
- For this course 02 assignments are uploaded on the AIOU Aaghi Portal along with allied material. You are advised to complete your assignments within the required time and upload it to your assigned tutor.
- This is compulsory course work, and its successful completion will make you eligible to take the final examination at the end of the semester.
- You will upload your assignments to your appointed tutor, whose name is notified to you for assessment and necessary guidance through concerned Regional Office of AIOU. You can also locate your tutor through AIOU website. Your tutor will return your online assignments after marking and providing necessary academic guidance and supervision.

Workshops

- The online mandatory workshops through (LMS) Aghi Portal of Bachelor Studies BS Economics courses will be arranged during each semester or as-per AIOU policy. Attendance and course quizzes are compulsory in workshops. A student will not be declared pass until he/she attends the workshop satisfactorily and actively.
- The duration of a workshop for each 03-credit course will be as per AIOU policy.

Revision before the Final Examination

It is very important that you revise the course as systematically as you have been studying.

You may find the following suggestions helpful.

- Go through the course unit one by one, using your notes during tutorial meetings to remind you of the key concepts or theories. If you have not already made notes, do so now.
- Prepare a chronology with short notes on the topics/events/personalities included in all units.
- Go through your assignments and check your weak areas in each case.
- Test yourself on each of the main topics, write down the main points or go through all the notes.
- Make sure to attend the workshops and revise all the points that you find difficult to comprehend.
- Try to prepare various questions with your fellow students during the last few tutorial meetings. A group activity in this regard is helpful. Each student should be given a topic and revise his topics intensively, summarize it and revise in group, then all members raise queries and questions. This approach will make your studies interesting and provide you with an opportunity to revise thoroughly.
- For the final exam paper, go through last semesters' papers. This can clarify questions and decide how to frame an answer.
- Before your final exams, make sure that,
  - you get your roll-number slip.
  - you know the exact location of the examination center.
  - You know the date and time of the examination.

Note:
This study guide has been developed to guide the students about the course "Fundamentals of Econometrics". In this context we want to make it clear that you are not bound to depend entirely upon the recommended books in the study guide. In case you are unable to find any recommended book, please feel free to consult any other book which covers the main contents of the course.

Moreover, you can get information regarding your Assignments, Workshop Schedule, Assignment, Results, Tutors, and Final Examination from the AIOU website: www.aiou.edu.pk and through your LMS account. You are advised to regularly visit the university website to update yourself about the activities.

COURSE OUTLINE

This course is comprised of the following units.

UNIT 01: An Introduction to Econometrics:

- Why Study Econometrics?
- What is Econometrics?
- Economic and Econometric Model
- Nature and Sources of Data for Econometric Analysis

UNIT 02: Two Variables Linear Regression Model:

- Introduction
- The Concept of the Population Regression Function (PRF)
- The Significance of the Stochastic Disturbance Term
- The method of Ordinary Least Squares (OLS)
- Assumptions of The Ordinary Least Squares Method
- Properties of The Least Squares Method Measures of the Goodness of Fit
- The Probability Distribution of Disturbance Term
- The Normality assumption on Disturbance Term
- Properties of OLS Estimator under Normality Assumptions
- The Method of Maximum Likelihood (ML)
- Statistical Inference in the Linear Regression Model
- Analysis of Variance of the Linear Regression Model

UNIT 03: Multiple Regression Models:

- Introduction
- A Model with Two Explanatory Variables
- Statistical Inference in The Multiple Regression Model
- Interpretation of the Regression Coefficients
- Partial and Multiple Correlation coefficients and their Relationship
- Prediction in the Multiple Regression Model
- The Multiple Coefficient of Determination
- Analysis of Variance and Tests of Hypothesis
- Tests for Stability

UNIT 04: The Matrix Approach to Linear Regression Model

- Introduction
- The k-Variable Linear Regression Model
- Assumptions of Linear Regression model in Matrix Notations

- OLS Estimation and Properties of OLS Estimators
- Hypothesis Testing in Matrix Notations
- Analysis of Variance in Matrix Notation
- The Correlation Matrix

UNIT 05: Multicollinearity:

- Introduction
- Nature of the Multicollinearity
- Estimation in the Presence of Multicollinearity
- Consequences of Multicollinearity
- Detection of Multicollinearity
- Remedial Measures

UNIT 06: Heteroscedasticity:

- Introduction
- Nature of the Heteroscedasticity
- Detection of Heteroscedasticity
- Consequences of Heteroscedasticity
- Solutions to Heteroscedasticity Problems

UNIT 07: Autocorrelation:

- Introduction
- Nature of The Autocorrelation
- Consequences of Autocorrelation
- Methods of Detection of Autocorrelation
- Remedial Measures

UNIT 08: Model Specification and Diagnostic Testing

- Introduction
- Model Selection Criteria
- Types of Specification Errors
- Consequences of Model Specification Errors
- Tests of Specification Errors
- Errors of Measurements
- Nested Versus Non-Nested Models
- Tests of Non-Nested Hypothesis
- Model Selection Criteria in Nested and Non-Nested Models

UNIT 09: Simultaneous Equation Models:

- Introduction
- The Nature of the Simultaneous Equation Models
- Endogenous and Exogenous Variables
- Structural Equations and Reduced form Equations
- The Identification Problem
- Methods of Identification
- Methods of Estimations (OLS, ILS, 2SLS)
- Limitations of Dynamic Analysis

Textbooks & Supplies:

1. Gujrati, D. J. - Basic Econometrics (Latest Edition) McGraw-Hill Company.
2. Maddala, G. S. – Econometrics (Latest Edition) – McGraw-Hill Company.
3. Koutsoyiannis, A.- Theory of Econometrics (Latest Edition) - McMillan.

Additional Readings:

1. Dougherty, Christopher – Introduction to Econometrics (Latest edition) Oxford University Press.
2. Free online course on Introduction to Econometrics, available from http://asadzaman.net.
3. Pindyck & Rubinfeld- Econometric Models & Economic Forecasts (Latest Edition) McGraw-Hill Inc.
4. Stock H. J. and M. W. Watson, Introduction to Econometrics, India: Pearson Education.
5. Stewart G. K., Introduction to Applied Econometrics, United States of America: Curt Hinrichs.
6. Wonnacot & Wonnacot Econometrics (Latest Edition) -John Wiley, New York.

UNIT 01

AN INTRODUCTION TO ECONOMETRICS

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

1.1. INTRODUCTION

Econometrics, a discipline that marries economics, statistics, and mathematics, is a pivotal field of study for anyone seeking to understand the complex dynamics of economic systems. It provides the tools necessary to test economic theories, evaluate policy interventions, and make informed predictions based on real-world data. The study of econometrics equips students with the skills to critically analyze economic phenomena and contribute to evidence-based decision-making.

At the heart of econometrics are economic and econometric models. Economic models, grounded in economic theories and assumptions, offer a theoretical framework to understand how different economic agents interact in markets. Econometric models, on the other hand, enhance these economic models by integrating statistical techniques and real-world data. They serve as a bridge between theory and empirical analysis, providing a framework for quantitative assessment and hypothesis testing.

Data is the lifeblood of econometric analysis. Economists rely on different types of data, including time series, cross-sectional, and panel data, to study various aspects of economic behavior. The nature and sources of data for econometric analysis are diverse, ranging from government agencies and international organizations to surveys and field experiments. The quality and availability of data are crucial for ensuring accurate and meaningful econometric results (Gujarati & Porter, 2009; Sharma, 2023).

In conclusion, the study of econometrics provides a systematic framework for analyzing economic relationships and making predictions. By combining economic theory with statistical methods, economists can extract valuable insights from economic data and contribute to the empirical understanding of economic phenomena.

1.2. OBJECTIVES

While delving into the sections mentioned, students should aim to achieve the following objectives:

- **understanding the importance of Econometrics:** Students should comprehend why econometrics is a crucial field of study. They should understand how econometrics allows for the testing of economic theories, the evaluation of policy interventions, and the making of informed predictions based on real-world data.

- **grasping the Concept of Econometrics:** Students should be able to define econometrics and explain how it combines economics, statistics, and mathematics to analyze economic data.
- **differentiating between Economic and Econometric Models:** Students should understand the distinction between economic and econometric models. They should recognize how econometric models enhance economic models by integrating statistical techniques and real-world data.
- **recognizing the Role of Data in Econometrics:** Students should appreciate the importance of data in econometric analysis. They should understand the different types of data (time series, cross-sectional, and panel data) and the various sources from which this data can be obtained.
- **applying Econometric Principles:** Students should aim to apply the principles and techniques of econometrics in analyzing economic phenomena. This includes understanding how to use econometric models to test economic theories, evaluate policy interventions, and make informed predictions.
- **critical Thinking and Analysis:** Students should develop the ability to critically analyze economic phenomena using econometric tools. This includes questioning assumptions, interpreting results, and understanding the limitations of econometric analysis.
- **connecting Theory with Practice:** Students should strive to connect theoretical concepts with practical applications. This includes understanding how econometric models can be used to analyze real-world economic issues and inform policy decisions.

1.3. Major Topics

Following are the major topics that are discussed with detail in section 1.4.

- Why study Econometrics?
- What is Econometrics?
- Economic and Econometric Model
- Nature and Sources of Data for Econometric Analysis

1.4. Summary of the Unit

1.4.1. The Significance of Studying Econometrics

Econometrics is a vital field of study that bridges the gap between economic theory and empirical analysis. It equips economists with the necessary tools to rigorously test economic theories, evaluate policy interventions, and make informed predictions based on real-world data. By studying econometrics, students gain the skills required to critically analyze economic phenomena and contribute to evidence-based decision-making in both academia and policy circles.

The importance of econometrics in economic analysis cannot be overstated. It provides a robust framework for testing economic theories, forecasting future trends, and informing policy decisions. As the study by Koutsoyiannis (1977) highlighted, econometrics is a crucial tool for economists as it allows them to quantify relationships between economic variables, thereby enabling them to make predictions about future economic conditions based on past data.

Take, for example, the theory of the income-consumption relationship. This theory posits that an increase in income should lead to a corresponding increase in consumption expenditure. Econometric analysis enables economists to collect and analyze data on income and consumption expenditure from households. By employing statistical techniques, economists can estimate the relationship between these variables and assess its significance and magnitude. This empirical analysis provides insights into the actual behavior of individuals and households, enriching our understanding of economic phenomena.

The value of studying econometrics is rooted in its capacity to provide empirical evidence that either supports or refutes economic theories. Economic theory provides valuable insights into human behavior and market dynamics, but it often relies on simplifying assumptions that may not fully encapsulate the complexity of real-world economic interactions. Econometric analysis allows economists to test these theoretical assumptions using real-world data, thereby gaining a deeper understanding of economic behavior and drawing robust conclusions.

Moreover, econometrics is not only useful for economists but also for various other fields. For instance, in finance, econometric models are used to predict stock prices, interest rates, and other financial variables. In the public sector, econometric analysis is used to evaluate the effectiveness of policy measures and to forecast the impact of proposed policies.

In the context of economic development, econometrics plays a pivotal role. The studies emphasize the role of econometrics in understanding the complex dynamics of economic growth. By using econometric techniques, researchers can identify the key drivers of economic growth and assess the impact of various factors such as investment, education, and technological progress on economic development.

However, it's important to note that while econometrics is a powerful tool, it is not without its limitations. As pointed out by Leamer (1983), econometric models are based on assumptions, and if these assumptions are not met, the results can be misleading. Therefore, it is crucial to use econometric methods with caution and to interpret the results in the light of the underlying assumptions and the quality of the data used.

Based on the book by Gujarati and Porter (2009), the importance of studying econometrics can be understood from several perspectives:

- Empirical Testing of Economic Theories: Econometrics provides tools for testing economic theories. For instance, the theory of demand suggests that there is a negative relationship between price and quantity demanded. Econometrics allows us to test this theory using real-world data (Gujarati & Porter, 2009, p. 5).
- Forecasting: Econometrics can be used to forecast economic variables. For example, econometric models can be used to predict future GDP, inflation rates, stock prices, and so on. These forecasts can be invaluable for policymakers and businesses (Gujarati & Porter, 2009, p. 7).
- Policy Evaluation: Econometrics is also crucial for policy evaluation. Policymakers often need to know the potential impact of their policies before they are implemented. Econometric models can be used to simulate the effects of various policy proposals, helping policymakers make informed decisions (Gujarati & Porter, 2009, p. 9).
- Quantitative Analysis of Economic Phenomena: Econometrics allows economists to quantify economic phenomena. For example, it can be used to estimate the elasticity of demand for a product, the impact of education on wages, and so on. This quantitative analysis can provide valuable insights that are not immediately apparent from a qualitative analysis (Gujarati & Porter, 2009, p. 11).

In conclusion, econometrics is an indispensable tool in economic analysis. It provides a rigorous and systematic approach to understanding economic phenomena and making predictions about future economic conditions. However, like any tool, it must be used with care and understanding of its limitations.

1.4.2. What is Econometrics?

Econometrics, a term coined by Ragnar Frisch (1933), is a multidisciplinary field that marries economics, statistics, and mathematics. It is a scientific discipline that applies statistical and mathematical methods to the analysis and interpretation of economic data, thereby enabling the empirical verification of economic theory and the quantification of economic phenomena (Gujarati, 2011).

At the heart of econometrics is the development and application of econometric models. These models, often expressed in mathematical form, provide a structured framework for quantifying and exploring the relationships between different economic variables. The most fundamental of these is the linear regression model, which posits a linear relationship between a dependent variable and one or more independent variables (Gujarati, 2011).

The process of estimating econometric models involves the application of statistical techniques to estimate unknown parameters, such as regression coefficients, based on observed data. A commonly employed method is the Ordinary Least Squares (OLS) estimation, which minimizes the sum of squared residuals between the observed and predicted values of the dependent variable (Gujarati, 2011). This method has been widely used in econometrics due to its desirable statistical properties, such as unbiasedness and efficiency under certain conditions (Wooldridge, 2012).

Once the parameters of the model have been estimated, the next step involves the analysis of the statistical significance and economic interpretation of the estimated coefficients. Statistical hypothesis tests, such as t-tests and F-tests, are conducted to determine the significance of the relationships between variables (Gujarati, 2011). The economic interpretation of the coefficients provides insights into the direction and magnitude of the relationships, allowing economists to draw meaningful conclusions about economic phenomena (Wooldridge, 2012).

Econometric models also serve as a basis for making predictions about future economic outcomes. By utilizing the estimated model parameters, economists can forecast the values of the dependent variable under different scenarios or policy changes. These predictions provide valuable insights for understanding the potential impact of economic factors and policy interventions on economic outcomes (Stock & Watson, 2015).

However, econometrics is not without its challenges and limitations. Econometric analysis relies on several assumptions, such as linearity, independence, and homoscedasticity, which need to be carefully considered and tested. Violations of these assumptions can lead to biased or inefficient estimates (Gujarati, 2011). Additionally, issues such as endogeneity, measurement errors, and omitted variable bias can pose challenges in econometric analysis. Econometricians employ advanced techniques, such as instrumental variable estimation and panel data methods, to address these challenges and ensure the robustness and validity of their findings (Wooldridge, 2012).

In conclusion, econometrics offers a systematic framework for analyzing economic relationships and making predictions. By combining economic theory with statistical methods, economists can extract valuable insights from economic data and contribute to the empirical understanding of economic phenomena. The field of econometrics continues to evolve, with ongoing research and development of new methods and techniques to address the complex challenges of economic analysis (Stock & Watson, 2015).

1.4.3. Economic and Econometric Model

The field of economics utilizes models as powerful tools to simplify complex economic systems and elucidate the relationships between different economic variables. Economic models are constructed based on economic theories and assumptions, providing a framework to comprehend how individuals, households, firms, and governments make economic decisions and interact in markets. These models offer theoretical insights into economic behavior and enable predictions regarding the impact of changes in economic variables.

However, economic models often incorporate simplifying assumptions to focus on specific aspects of economic behavior or facilitate analysis. While these assumptions allow economists to derive theoretical insights, they may not capture the full complexity of real-world economic phenomena. This is where econometric models play a crucial role. Econometric models advance economic models by integrating statistical techniques and real-world data. They serve as a bridge between theory and empirical analysis, providing a framework for quantitative assessment and hypothesis testing.

Econometric models often specify the mathematical form that represents the relationship between economic variables. For instance, a simple linear regression model is employed to study the relationship between two variables, Y and X, and can be expressed as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1.1)$$

Here in equation 1.1, Y represents the dependent variable, X signifies the independent variable, β_0 and β_1 denote the regression coefficients, and ε captures the error term accounting for unobserved factors influencing Y. The estimation of coefficients β_0 and β_1 employs econometric techniques to quantify the relationship between Y and X based on available data.

Econometric models enable economists to test economic theories against real-world data. By estimating model parameters and conducting hypothesis tests, economists can evaluate the statistical significance of the relationships proposed by economic theory. This empirical analysis provides evidence to support or reject economic hypotheses, leading to refinements of economic models to better align with real-world behavior.

Moreover, econometric models facilitate policy simulations and counterfactual analyses. By manipulating the values of independent variables within the model, economists can assess the potential impact of policy interventions or changes in economic variables on the dependent variable. This aids policymakers in understanding the potential consequences of various policy choices and designing more effective economic policies.

Econometric models are not confined to linear relationships or simple models. They can encompass more intricate relationships, non-linearities, and interactions among variables. Advanced econometric techniques, such as panel data analysis, time series analysis, and simultaneous equation models, enable economists to explore more complex economic phenomena and capture the dynamic nature of economic relationships.

Panel data analysis, as explained by Gujarati (2003), is a method that combines cross-sectional data (data collected at one point in time across several subjects) and time-series data (data collected over time for a single subject) to provide a richer dataset. This method allows for the control of individual heterogeneity, improves the efficiency of econometric estimates, and enables the modeling of complex behavioral patterns that cannot be captured by cross-sectional or time-series data alone (Hsiao, 2007).

Time series analysis, on the other hand, focuses on data collected over time for a single subject. It is particularly useful in forecasting and understanding the underlying forces and structure that produced the observed data. Gujarati (2003) explains that time series data may involve trends, seasonality, cycles, and irregular movements. Econometric models using time series data can help in understanding and predicting the behavior of these variables over time. For instance, Box and Jenkins (1970) developed the autoregressive integrated moving average (ARIMA) model, a cornerstone in time series forecasting.

Simultaneous equation models, as the name suggests, involve a system of equations where the endogenous variables are determined simultaneously. These models are particularly useful in situations where there are circular causal relationships between variables, making it difficult to distinguish between dependent and independent variables (Gujarati, 2003). The simultaneous equation models have been extensively used in macroeconomics and econometrics to understand the complex interrelationships between economic variables (Bowden & Turkington, 1990).

In summary, econometric models build upon economic models by incorporating statistical techniques and real-world data. They offer a systematic framework for estimating and testing economic relationships, evaluating economic theories, conducting policy analyses, and making predictions. Econometric modeling enhances our understanding of economic behavior and facilitates evidence-based decision-making. The use of advanced econometric techniques, such as panel data analysis, time series analysis, and simultaneous equation models, allows for a more comprehensive and nuanced understanding of economic phenomena.

1.4.4. Nature and Sources of Data for Econometric Analysis

In the realm of econometric analysis, data is the bedrock upon which all investigations and estimations are built. The quality, relevance, and representativeness of the data used can significantly influence the accuracy and reliability of econometric models. Economists use different types of data, including time series, cross-sectional, and panel data, to study various aspects of economic behavior and apply specific econometric techniques for analysis.

Time series data, which consists of observations of a variable over time, is a common type of data used in econometric analysis. This type of data captures the dynamics and trends of economic variables over a specific time period and can include annual GDP growth rates, monthly unemployment rates, or daily stock prices. Time series data allows economists to explore interdependencies and patterns in economic variables over time, providing valuable insights into the temporal dynamics of economic phenomena (Gujarati, 2003).

Cross-sectional data, on the other hand, consists of observations collected at a specific point in time, typically representing different individuals, regions, or countries. This type of data can include household surveys, industry-level data, or census data. Cross-sectional data provides insights into the differences and variations in economic variables across distinct units or groups. It allows economists to examine the heterogeneity among different units and understand the variations in economic behavior across different entities (Wooldridge, 2012).

Panel data combines elements of both time series and cross-sectional data. It consists of observations on multiple variables for a specific set of units, such as

individuals, firms, or countries, over time. Panel data provides valuable insights into individual dynamics and facilitates the analysis of both within-unit and between-unit variations. Panel data analysis is particularly beneficial for studying individual behavior, evaluating policy impacts, and capturing the effects of time-varying variables (Baltagi, 2008).

Economists obtain data from diverse sources to conduct econometric analysis. Government agencies, central banks, statistical offices, and international organizations serve as primary sources of economic data. These institutions collect and publish data on macroeconomic variables, labor markets, trade, inflation, and other economic indicators. Additionally, economists may undertake their data collection exercises, such as surveys or field experiments, to gather specific data pertinent to their research questions (Beck & Katz, 1995).

However, data often requires preprocessing and transformation before it is suitable for econometric analysis. Missing data, outliers, or measurement errors necessitate addressing through imputation, robust estimation techniques, or data cleaning procedures. Econometricians must also consider issues such as endogeneity, sample selection bias, and measurement errors that can impact the estimation and interpretation of econometric models (Gujarati & Porter, 2009).

In conclusion, the quality and availability of data are essential for econometric analysis. Different types of data, including time series, cross-sectional, and panel data, provide insights into different aspects of economic behavior and require specific econometric techniques. Reliable and relevant data from trusted sources are crucial to ensuring accurate and meaningful econometric results.

1.5. Self-Assessment Questions

- What is econometrics and why is it important to study it?
- What is the difference between an economic model and an econometric model? Provide examples.
- What are the different types of data used in econometric analysis and what insights do they provide?
- What are some of the sources from which economists obtain data for econometric analysis?
- What are some of the challenges and limitations associated with econometric analysis?
- How do econometric models contribute to policy simulations and counterfactual analyses? Provide an example.

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. 2nd Edition, McMillan.

Additional Readings

- Angrist, J. D., & Pischke, J. S. Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Baltagi, B. H. Econometric analysis of panel data. John Wiley & Sons.
- Beck, N., & Katz, J. N. (1995). What to do (and not to do) with Time-Series Cross-Section Data. American Political Science Review, 89(3), 634-647.
- Bowden, R. J., & Turkington, D. A. Instrumental variables. Cambridge University Press.
- Box, G. E., & Jenkins, G. M. Time series analysis: Forecasting and control. Holden-Day.
- Cameron, A. C., & Trivedi, P. K. Micro-econometrics: methods and applications. Cambridge university press.
- Frisch, R. Pitfalls in the statistical construction of demand and supply curves. Verlag von Julius Springer.
- Greene, W. H. Econometric analysis. Pearson Education India.
- Gujarati, D. N. Basic Econometrics, Latest edition, McGraw-Hill.
- Gujarati, D. N. Econometrics by Example. Palgrave Macmillan.
- Hayashi, F. Econometrics. Princeton University Press.
- Hsiao, C. (2007). Panel data analysis—advantages and challenges. Test, 16(1), 1-22.
- Kennedy, P. A guide to econometrics. John Wiley & Sons.
- Leamer, E. E. (1983). Let's take the con out of Econometrics. *The American Economic Review*, 73(1), 31-43.
- Stock, J. H., & Watson, M. W. Introduction to Econometrics. Pearson.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 02

TWO VARIABLES LINEAR REGRESSION MODEL

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

2.1. INTRODUCTION

In the study of econometrics, we encounter important concepts in regression analysis. The Population Regression Function (PRF) forms the basis for understanding relationships between variables. Stochastic disturbance terms represent unseen factors affecting the model, adding uncertainty.

Ordinary Least Squares (OLS) is a popular method to estimate regression coefficients, offering reliable results. To ensure OLS's effectiveness, we must consider its assumptions, which guarantee consistent estimations. OLS properties make it efficient and powerful compared to other unbiased methods. Evaluating the model's fit, Measures of Goodness of Fit assess how well the model captures data variability. Additionally, considering the Probability Distribution of Disturbance Term helps us understand error assumptions and enables robust statistical analysis. Exploring normality assumptions further enhances the statistical properties of OLS estimators, facilitating hypothesis testing. For more insights into parameter estimation, we may also consider the Method of Maximum Likelihood (ML), which aligns with OLS under certain conditions.

By delving into these concepts, researchers gain valuable knowledge and skills to navigate the complexities of regression analysis and draw meaningful conclusions from data.

2.2. OBJECTIVES

While reading the sections of this document, students should aim to achieve the following objectives:

- understand the concept of the Population Regression Function (PRF) and its significance in econometrics.
- comprehend the significance of the Stochastic Disturbance Term in the regression model.
- learn about the method of Ordinary Least Squares (OLS) and its application in estimating regression coefficients.
- familiarize themselves with the assumptions of the Ordinary Least Squares Method and understand their importance in ensuring the effectiveness of OLS.
- understand the properties of the Least Squares Method and why it is preferred over other methods.
- learn about the measures of the Goodness of Fit and how they assess the model's performance.

- understand the Probability Distribution of the Disturbance Term and its role in statistical analysis.
- learn about the Normality assumption on the Disturbance Term and its importance in enhancing the statistical properties of OLS estimators.
- understand the properties of OLS Estimator under Normality Assumptions.
- learn about the Method of Maximum Likelihood (ML) and its application in parameter estimation.
- understand the concept of Statistical Inference in the Linear Regression Model and its importance in drawing conclusions from the model.
- learn about the Analysis of Variance of the Linear Regression Model and its role in understanding the variability in the data.

By achieving these objectives, students will gain a comprehensive understanding of the key concepts in econometrics and be able to apply these concepts in practical scenarios.

2.3. Major Topics

- The Concept of the Population Regression Function (PRF)
- The Significance of the Stochastic Disturbance Term
- The method of Ordinary Least Squares (OLS)
- Assumptions of The Ordinary Least Squares Method
- Properties of The Least Squares Method
- Measures of the Goodness of Fit
- The Probability Distribution of Disturbance Term
- The Normality assumption on Disturbance Term
- Properties of OLS Estimator under Normality Assumptions
- The Method of Maximum Likelihood (ML)
- Statistical Inference in the Linear Regression Model
- Analysis of Variance of the Linear Regression Model

2.4. Summary of the Units

2.4.1. The Concept of the Population Regression Function (PRF)

The Population Regression Function (PRF) is a fundamental concept in the field of econometrics and statistics. It represents the relationship between a dependent variable and one or more independent variables in a population. The PRF is typically expressed in the form of an equation, which can be linear or non-linear depending on the nature of the relationship between the variables.

The PRF, in essence, is an equation that delineates the average value of the dependent variable as a function of the explanatory variables. It is important to note that this equation is not directly observable, but rather, it is an underlying truth that we strive to approximate through statistical methods. The PRF is often represented as $E(Y|X)$, where E denotes the expected value, Y is the dependent variable, and X represents the explanatory variables. This notation underscores the conditional nature of the PRF, indicating that the expected value of Y is contingent upon the values of X.

In the context of linear regression, the PRF is typically assumed to be a linear function. This assumption, while not always accurate, simplifies the analysis and provides a useful starting point for understanding the relationship between variables. However, it is crucial to remember that the true PRF may be nonlinear, and further investigation may be necessary to accurately capture the complexity of

the relationship. In its simplest form, the PRF for a linear regression model with one independent variable is expressed as:

$$Y = \beta_0 + \beta_1 X + u \quad (2.1)$$

where: Y is the dependent variable, X is the independent variable, β_0 and β_1 are parameters of the model, and u is the error term.

The parameters β_0 and β_1 represent the intercept and slope of the regression line, respectively. The intercept (β_0) is the value of Y when X is zero, and the slope (β_1) represents the change in Y for a one-unit change in X. The error term (u) captures the influence of all other factors not included in the model that affect the dependent variable. It is assumed to have a mean of zero and is uncorrelated with the independent variable.

The PRF provides a theoretical framework for understanding the relationship between variables in a population. However, in practice, we usually do not have access to the entire population data and have to estimate the PRF using a sample. This leads to the concept of the Sample Regression Function (SRF), which is an estimate of the PRF based on sample data. It's important to note that the PRF is a deterministic function, meaning it provides a fixed output for a given input. On the other hand, the SRF is a stochastic function, meaning it includes a random error term to account for the variability in the data.

2.4.2. The Significance of the Stochastic Disturbance Term

In the realm of econometric analysis, the stochastic disturbance term, often denoted as ' u ', holds a position of paramount importance. It is a term that encapsulates the myriad of factors that are not explicitly included in the model but nonetheless exert an influence on the dependent variable (Greene, 2003). The stochastic disturbance term is a random variable with a mean of zero, and it is uncorrelated with the explanatory variables. This assumption is crucial for the Ordinary Least Squares (OLS) estimator to be unbiased and consistent (Wooldridge, 2012).

The equation for a simple linear regression model, inclusive of the stochastic disturbance term is presented in 2.1, where u is the stochastic disturbance term. The term u is assumed to have a normal distribution with a mean of zero and a constant variance, σ^2. This assumption, known as homoscedasticity, is vital for the efficient operation of the OLS estimator (Gujarati, 2003).

The stochastic disturbance term, u, is a catch-all for all the unobserved factors that affect the dependent variable Y but are not included as explanatory variables in the model. These factors could include measurement errors, omitted variables, or

random shocks. The inclusion of the stochastic disturbance term in the model allows us to capture these unobserved effects and thus provides a more realistic and comprehensive representation of the relationship between the dependent and explanatory variables (Stock & Watson, 2011).

In conclusion, the stochastic disturbance term is a fundamental component of econometric models. It encapsulates the unobserved factors influencing the dependent variable, allowing for a more comprehensive and realistic representation of the relationship under study. Its assumptions are critical for the unbiasedness and consistency of the OLS estimator, making it a vital element in econometric analysis.

2.4.3. The Method of Ordinary Least Squares (OLS)

The method of Ordinary Least Squares (OLS) is a widely used statistical technique employed in econometrics to estimate the parameters of a linear regression model. It is a powerful tool for analyzing the relationship between a dependent variable and one or more explanatory variables.

In the context of a linear regression model with one explanatory variable, the OLS method aims to find the line that best fits the observed data points. The goal is to minimize the sum of squared differences (residuals) between the actual values of the dependent variable and the predicted values based on the linear relationship with the explanatory variable.

Let's consider again equation (2.1), where Y represents the dependent variable (the variable we want to predict or explain), X is the explanatory variable (the variable we believe influences Y), β_0 is the intercept of the regression line (the value of Y when X is zero), β_1 is the slope of the regression line (the change in Y for a unit change in X), and u is the error term (representing the unexplained variation in Y that is not captured by the linear relationship with X).

The OLS method estimates the values of β_0 and β_1 that minimize the sum of squared residuals:

$$min \sum (Y_i - (\beta_0 + \beta_1 X_i))^2 \qquad (2.2)$$

where the summation is taken over all the observed data points ($i = 1 \; to \; n$), and Y_i and X_i are the actual values of the dependent and explanatory variables, respectively.

The OLS estimator provides the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that result in the best-fitting line through the data points, minimizing the vertical distance between the data

points and the regression line. It aims to find the line that best explains the relationship between X and Y based on the available data.

The formulas to estimate β_0 and β_1 using the OLS method are as follows:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad (2.3)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (2.4)$$

Where: Σ denotes the summation symbol, $\bar{X}$ is the mean of the X values, $\bar{Y}$ is the mean of the Y values.

The slope ($\hat{\beta}_1$) measures the change in Y for a one-unit change in X, accounting for the relationship between the two variables. The intercept ($\hat{\beta}_0$) represents the value of Y when X is zero (if applicable) and is derived by subtracting the product of the slope ($\hat{\beta}_1$) and the mean of X ($\bar{X}$) from the mean of Y ($\bar{Y}$).

So, OLS estimator of the slope ($\hat{\beta}_1$) can be obtained by the ratio of covariance between X and Y ($Cov(X,Y) = \sigma_{XY}$) divided by the variance of X ($Var(X) = \sigma_X^2$).

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_{XY}}{\sigma_X^2} \quad (2.5)$$

Or by using the formula of correlation between X and Y, $Corr(X,Y) = r_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$, we can write formula for OLS estimator of slope ($\hat{\beta}_1$) as follows:

$$\hat{\beta}_1 = \frac{Corr(X,Y)}{Var(X)} = \frac{r_{XY} \cdot \sigma_Y}{\sigma_X} \quad (2.6)$$

Likewise, OLS estimator of the slope ($\hat{\beta}_1$) presented in (2.3), can be written in the deviation form as follow:

$$\hat{\beta}_1 = \frac{\sum xy}{\sum x^2} \quad (2.7)$$

Where, $x = X_i - \bar{X}$ represents the deviation of variable X from its mean $\bar{X}$ and $y = Y_i - \bar{Y}$ represents the deviation of dependent variable Y from its mean $\bar{Y}$. There is one another formula which can be employed to get the OLS estimator calculates the slope ($\hat{\beta}_1$) which is presented as follow:

$$\hat{\beta}_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \quad (2.8)$$

Once both the OLS estimator slope ($\hat{\beta}_1$) and the intercept ($\hat{\beta}_0$) are estimated, they together define the linear regression model. This model helps economists understand the relationship between the variables and make predictions based on the linear relationship observed in the data. The OLS estimators play a vital role in econometrics by allowing us to interpret the strength and direction of the relationship between the variables and draw meaningful conclusions from the data.

In conclusion, the method of Ordinary Least Squares is a powerful tool to estimate the parameters of a linear regression model with one explanatory variable. By

minimizing the sum of squared residuals, the OLS method identifies the best-fitting line that explains the relationship between the variables. The estimated values of β_0 and β_1 allow economists to analyze the strength and direction of the relationship and make predictions based on the linear model.

2.4.4. Assumptions of the Ordinary Least Squares Method

The assumptions of Ordinary Least Squares (OLS) are:

2.4.4.1. Linearity in Parameters

This assumption posits that the regression model is linear in the parameters. In other words, the dependent variable (Y) is a linear function of the parameters ($\beta_0, \beta_1, \dots, \beta_k$) and the error term ($u$). It doesn't necessarily mean that the relationship between Y and the independent variables ($X_1, X_2, \dots, X_k$) is linear, but rather that the parameters are linear. This assumption allows us to use linear algebra to estimate the parameters.

2.4.4.2. Random Sampling

This assumption states that the observations are randomly drawn from the population. Each observation of the dependent and independent variables is a random draw from the underlying population. This is crucial for the generalizability of the results. In the context of time series data, this assumption usually implies that the time series is stationary. It is assumed that the X variable(s) and the error term are independent, that is:

$$Cov\ (X_i\ , u_i) \ = \ 0 \qquad (2.9)$$

2.4.4.3. Zero Conditional Mean of Disturbance

This assumption implies that the error term (u) has a zero population mean given any value of the explanatory variables. In other words, the errors, on average, cancel out. This assumption is crucial for the OLS estimators to be unbiased. Given the value of X_i, the mean, or expected, value of the random disturbance term u_i is zero. Symbolically, we have:

$$E\ (u_i|X_i) \ = \ 0 \qquad (2.10)$$

Or, if X is non-stochastic,

$$E\ (u_i) \ = \ 0 \qquad (2.11)$$

2.4.4.4. Homoscedasticity or Constant Variance of Disturbance

This assumption states that the error term has the same variance given any value of the explanatory variables. This means that the spread or dispersion of the distribution of the error term does not change across different levels of the

independent variables. In other words, the variance of the error, or disturbance, term is the same regardless of the value of X. Symbolically,

$$Var\ (u_i) = E[u_i - E(u_i|X_i)]^2$$
$$= E(u_i^2|X_i) \quad \text{because of assumption 2.4.4.3}$$
$$= E(u_i^2) \quad \text{if } X_i \text{ is non-stochastic}$$
$$= \sigma^2 \qquad (2.12)$$

Where, Var stands for variance.

If this assumption is violated (i.e., if the error variance changes across different levels of the independent variables, a situation known as heteroscedasticity), it can lead to inefficient and potentially biased estimates.

$$Var\ (u_i) \neq \sigma^2 \quad \text{or} \quad Var\ (u_i) = \sigma_i^2 \qquad (2.13)$$

2.4.4.5. No Autocorrelation between the Disturbances

This assumption asserts that the error term of one observation is not correlated with the error term of any other observation. Given any two X values, X_i and X_j $(i \neq j)$, the correlation between any two u_i and u_j $(i \neq j)$ is zero. In short, the observations are sampled independently. Symbolically,

$$Cov\ (u_i, u_j|X_i, X_j) = 0$$
$$Cov\ (u_i, u_j) = 0 \quad \text{if } X_i \text{ is non-stochastic} \qquad (2.14)$$

Where i and j are two different observations and where Cov means covariance. If this assumption is violated (i.e., if there is autocorrelation), it can lead to inefficient estimates and can also affect the validity of standard hypothesis tests.

$$Cov\ (u_i, u_j) \neq 0 \quad \text{or} \quad Cov\ (u_i, u_j) = \rho \qquad (2.15)$$

Where ρ is the coefficient of autocorrelation.

2.4.4.6. Number of Observations

The Number of observations (n) must be greater than the number of parameters (k) to be estimated. Alternatively, the number of observations must be greater than the number of explanatory variables.

2.4.4.7. Nature of X Variables

The X values in each sample must not all be the same. Technically, $Var\ (X)$ must be a positive number. Furthermore, there can be no outliers in the values of the X variable, that is, values that are very large in relation to the rest of the observations.

2.4.4.8. Normality of Errors

This assumption posits that the error term is normally distributed. This assumption is especially important for hypothesis testing. If the errors are normally distributed,

it allows us to make statements about the probability distribution of the OLS estimators and conduct hypothesis tests.

Each of these assumptions is crucial for the OLS estimator to have desirable properties such as unbiasedness, efficiency, and consistency. Violations of these assumptions can lead to issues such as biased or inefficient estimates and can also affect the validity of hypothesis tests.

2.4.5. Properties of the Least Squares Estimators

The properties of OLS estimators are often referred to as the Gauss-Markov theorem, which establishes, under certain assumptions (e.g., homoscedasticity, no perfect multicollinearity), the OLS estimators satisfy Best Linear Unbiased Estimators (BLUE) properties. This theorem mathematically proves that OLS is the best among all unbiased linear estimators in terms of minimum variance. Here are the key properties of OLS estimators with reference to BLUE:

2.4.5.1. Best

The OLS estimators are the "best" among all linear unbiased estimators. This means that if we consider all possible unbiased linear estimators for the regression coefficients, the OLS estimators have the smallest variance, making them the most efficient. In other words, no other linear unbiased estimator can provide more precise and accurate estimates of the regression coefficients than the OLS estimators.

2.4.5.2. Linear

The OLS estimators are "linear" because they are obtained by taking linear combinations of the observed dependent variable and explanatory variables. The linear nature of OLS makes it computationally straightforward and allows for closed-form solutions.

2.4.5.3. Unbiased

The OLS estimators are "unbiased" because, on average, they provide estimates that are centered around the true population values of the regression coefficients. This property holds under the assumption that the errors have a mean of zero and are uncorrelated with the explanatory variables. This can be mathematically represented as:

$$E\left(\hat{\beta}\right) = \beta \qquad (2.17)$$

2.4.5.4. Estimators

The term "Estimators" refers to the fact that OLS provides a method to estimate the unknown parameters (regression coefficients) of the linear regression model based on the observed data.

These properties demonstrate the favorable characteristics of the OLS estimators, making them a widely used and reliable technique in econometrics. However, it is essential to verify that the assumptions of the linear regression model are met to ensure the validity of the OLS estimates and their interpretations.

2.4.6. The Coefficient of Determination R^2: A Measure of "Goodness of Fit"

The Coefficient of Determination, often denoted as R^2, is a statistical measure used in regression analysis to assess the "goodness of fit" of the regression model. It quantifies the proportion of the total variation in the dependent variable that is explained by the independent variables in the model. R^2 ranges from 0 to 1, where a higher value of R^2 indicates a better fit of the regression model to the data.
R^2 measures the strength and reliability of the relationship between the dependent variable (Y) and the explanatory variables ($X_1, X_2, \ldots, X_k$) in the regression model. It provides insights into how well the model captures the variation in the dependent variable and how much of that variation is due to the independent variables' influence.

The concept of R^2 was first introduced by American mathematician Karl Pearson in 1896 and was later developed further by Ronald Fisher in the 1910s. R^2 is widely used in various fields, including economics, social sciences, and engineering, to evaluate the effectiveness of regression models.

Mathematically, R^2 is defined as the ratio of the explained variation (sum of squares due to regression) to the total variation (total sum of squares):

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \qquad (2.18)$$

Where, ESS represents explained sum of square calculated by $\sum \hat{y}_i^2 = \sum\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2$ which is variation of Y values about their mean ($\bar{Y} = \bar{\hat{Y}}$), TSS represents the total sum of square calculated by $\sum y_i^2 = \sum(Y_i - \bar{Y})^2$.

The total variation in the observed Y values about their mean value can be partitioned into two parts, one explained sum of square (ESS) and the other unexplained or residual sum of square (USS). Mathematically,

$$TSS = ESS + USS \quad (2.19)$$

Diving both side of the equation by TSS, we obtain:

$$\frac{TSS}{TSS} = \frac{ESS}{TSS} + \frac{USS}{TSS}$$

$$1 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$$

$$1 = R^2 + \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$$

$$R^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum(Y_i - \bar{Y})^2} \quad (2.20)$$

Additionally, R^2 can be expressed in terms of the correlation coefficient (r) between the observed values of the dependent variable and the predicted values from the regression model:

$$R^2 = r^2$$

$$R^2 = \frac{n\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$$

$$R^2 = \frac{n\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

$$R^2 = \frac{n\sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2 - (\sum X_i)^2][n\sum Y_i^2 - (\sum Y_i)^2]}} \quad (2.21)$$

The Coefficient of Determination R^2 has several important properties that provide valuable insights into the goodness of fit of a regression model. These properties help in assessing the effectiveness and reliability of the model in explaining the variation in the dependent variable. Here are the key properties of R^2:

- R^2 ranges from 0 to 1. A value of R^2 equal to 0 indicates that the independent variables in the model do not explain any of the variation in the dependent variable. On the other hand, an R^2 value of 1 implies that the independent variables perfectly explain all the variation in the dependent variable.
- R^2 quantifies the proportion of the total variation in the dependent variable that is explained by the independent variables in the regression model. For example, if R^2 is 0.70, it means that 70% of the variation in the dependent variable is explained by the independent variables.
- R^2 is often referred to as a measure of "goodness of fit." It assesses how well the regression model fits the observed data points. A higher R^2 indicates a better fit, meaning that the model captures a larger portion of the variability in the dependent variable.
- R^2 is easy to interpret. It provides a single numerical value that summarizes the model's explanatory power. It allows researchers to communicate the

percentage of the variation in the dependent variable explained by the model to a broader audience.

- R^2 is directly related to the correlation coefficient (r) between the observed values of the dependent variable and the predicted values from the regression model. Specifically, R^2 is equal to the square of the correlation coefficient ($R^2 = r^2$).
- R^2 can be used to compare different regression models. Comparing R^2 values among alternative models helps in selecting the model that provides the best fit to the data and the highest explanatory power.

Other than the above mentioned properties, R^2 should be used carefully by keeping in mind following limitations:

- R^2 has limitations, especially in the context of multiple regression. It tends to increase with the addition of more explanatory variables, even if the new variables do not significantly improve the model's explanatory power. Adjusted R^2 is often used to address this issue.
  Adjusted $R^2 = \bar{R}^2 = 1 - \left[(1 - R^2)\left(\frac{n-1}{n-k}\right)\right]$ (2.22)
- R^2 may be inflated by overfitting, especially in complex models. Overfitting occurs when a model is too flexible and fits the noise in the data rather than the underlying true relationship. Adjusted R^2 penalizes for the number of variables in the model to address this issue.
- R^2 should not be interpreted as a test of causality. Even if R^2 is high, it does not imply a causal relationship between the independent and dependent variables. Causality requires additional empirical or experimental evidence.

In summary, R^2 is a valuable measure in regression analysis that quantifies the proportion of variation in the dependent variable explained by the independent variables. It provides an easy-to-understand summary of the model's goodness of fit and allows researchers to compare and select the best-fitting models. However, it is essential to interpret R^2 in conjunction with other diagnostic measures and be cautious about its limitations in certain situations.

2.4.7. The Probability Distribution of Disturbance Term

In the study of econometrics, we often want to understand the behavior of certain estimators, like $\hat{\beta}_1$. This estimator is a kind of average, calculated from our data, which includes both the values of our outcome variable (Y) and our predictor

variable (X). In our analysis, we usually consider these X values as fixed or non-random.

The value of $\hat{\beta}_1$ is ultimately a linear function of the random variable u_i, which is random by assumption. Therefore, the probability distribution of $\hat{\beta}_1$ (and also of $\hat{\beta}_0$) will depend on the assumption made about the probability distribution of u_i. This is crucial because we want to make guesses or inferences about the true population values of these estimators, and the nature of the probability distribution of u_i assumes an extremely important role in hypothesis testing.

However, the method we're using, Ordinary Least Squares (OLS), doesn't tell us anything about how u_i is distributed. This is a bit of a problem because it limits how much we can infer from our sample to the population. To get around this, we often assume that u_i follows a normal distribution. When we add this normality assumption to our existing assumptions, we get what's called the classical normal linear regression model (CNLRM). This model is more powerful because it allows us to make more precise inferences about the population and provides a stronger basis for hypothesis testing in regression analysis.

2.4.8. The Normality Assumption on Disturbance Term

The normality assumption of residuals is of utmost importance in Ordinary Least Squares (OLS) regression due to its impact on the validity and reliability of the statistical analysis. When the residuals, or errors, follow a normal distribution, the OLS estimators become unbiased, efficient, and attain the status of Best Linear Unbiased Estimators (BLUE). This property ensures that the estimated coefficients provide the most accurate representation of the underlying population parameters. Researchers heavily rely on these estimators to draw meaningful inferences about the relationships between variables and make precise predictions in various fields, such as economics, social sciences, and engineering.

The normality assumption plays a crucial role in hypothesis testing and constructing confidence intervals around the regression coefficients. These statistical tests depend on the assumption of normality to accurately assess the significance of explanatory variables and the overall goodness-of-fit of the model. Deviations from normality can lead to incorrect conclusions and undermine the validity of statistical inferences. By adhering to the normality assumption, researchers can conduct valid t-tests and F-tests, providing a strong basis for making well-founded decisions and understanding the significance of the explanatory variables in the regression model.

Furthermore, the normality assumption impacts the efficiency of OLS estimators, particularly in large samples. When errors follow a normal distribution, the OLS estimators converge to their true values at a faster rate, resulting in more accurate estimates as the sample size increases. This is essential for obtaining reliable estimates in empirical research, where large samples are often used. However, while OLS can be robust to moderate departures from normality in large samples, small samples or strong statistical inferences require careful consideration of the normality assumption. It is prudent for researchers to assess the normality assumption through various diagnostic tests and graphical methods to ensure the validity of their OLS results and maintain the integrity of their regression analysis.

2.4.9. Properties of OLS Estimator under Normality Assumptions

Under the normality assumptions, the Ordinary Least Squares (OLS) estimator in regression analysis possesses several important properties that make it a powerful and reliable tool for estimating population parameters:

2.4.9.1. Unbiasedness

When the errors in the model follow a normal distribution, the OLS estimator is unbiased. It means that on average, the OLS estimator provides estimates of the regression coefficients that are centered around the true population values. This property ensures that the OLS estimator is not systematically overestimating or underestimating the population parameters.

2.4.9.2. Efficiency

Among all unbiased estimators, the OLS estimator has the minimum variance under normality assumptions. It is the most efficient estimator, meaning it provides the most precise and reliable estimates of the population coefficients compared to other unbiased estimators. This efficiency is crucial in obtaining accurate and reliable estimates with smaller standard errors.

2.4.9.3. Best Linear Unbiased Estimator (BLUE)

The combination of unbiasedness and efficiency makes the OLS estimator the Best Linear Unbiased Estimator (BLUE) under normality assumptions. No other linear unbiased estimator can outperform the OLS estimator in terms of precision and accuracy.

2.4.9.4. Consistency

With normality assumptions, the OLS estimator is a consistent estimator. As the sample size increases, the OLS estimator converges to the true population parameters, providing more accurate estimates with larger sample sizes.

2.4.9.5. Normality of Residuals:

Under normality assumptions, the OLS residuals (or errors) are themselves normally distributed with mean zero. This property is beneficial for conducting valid statistical tests, constructing confidence intervals, and assessing model fit. $\hat{\beta}_0$ (being the linear function of u_i), is normally distributed with:

Mean: $E(\hat{\beta}_0) = \beta_0$

Variance: $Var(\hat{\beta}_0) = \sigma^2_{\hat{\beta}_0} = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$

In short: $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right)$ (2.23)

Similarly, $\hat{\beta}_1$ (being the linear function of u_i), is normally distributed with

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

Variance: $Var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum x_i^2}$

In short: $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$ (2.24)

2.4.9.6. Statistical Inference

The normality assumptions allow for valid statistical inference, such as hypothesis testing and confidence interval construction. This is because the OLS estimator's sampling distribution follows a normal distribution, enabling researchers to make robust statistical inferences about the regression coefficients. By the properties of the normal distribution, the variable Z, which is defined as:

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma^2_{\hat{\beta}_0}} \quad (2.25)$$

Where, Z follows the standard normal distribution, that is, a normal distribution with zero mean and unit (= 1) variance. Likewise, Z for the statistical inference about slope parameter can be written as:

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma^2_{\hat{\beta}_1}} \quad (2.26)$$

2.4.9.7. Independence of Estimators

The distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent from the distribution of $\hat{\sigma}^2$.

2.4.9.8. Best Unbiased Estimators (BUE)

In regression analysis, $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained through Ordinary Least Squares (OLS) have the minimum variance among all unbiased estimators, making them the most precise and reliable choice for estimating the population values, regardless of whether the relationship between variables is linear or not. This powerful finding by Rao (1965, p. 258) highlights the efficiency of OLS estimators, making them the best unbiased estimators in the entire class of estimation techniques. Therefore, we can say that the least-squares estimators are best unbiased estimators (BUE); that is, they have minimum variance in the entire class of unbiased estimators.

It is important to note that while the normality assumptions enhance the OLS estimator's properties, OLS can still provide consistent and valid estimates even when the normality assumption is violated, especially in large samples. However, adhering to the normality assumptions is crucial when making strong statistical inferences and ensuring the reliability of the regression results. Researchers should always assess the normality of residuals through diagnostic tests to evaluate the appropriateness of the OLS model for their data.

2.4.10. The Method of Maximum Likelihood (ML)

An alternative method of point estimation with stronger theoretical properties than Ordinary Least Squares (OLS) is the method of maximum likelihood (ML). The method of maximum likelihood, as the name indicates, consists in estimating the unknown parameters in such a manner that the probability of observing the given Y's is as high (or maximum) as possible.

When assuming that the error terms (u_i) are normally distributed, the ML and OLS estimators of regression coefficients (β's) are identical, holding true for both simple and multiple regressions. However, the ML estimator for the variance ($\hat{\sigma}^2$) is slightly biased compared to the unbiased OLS estimator. Nevertheless, as the sample size (n) increases, the two estimators of $\hat{\sigma}^2$ tend to become equal, making the ML estimator asymptotically unbiased. Since OLS, along with the assumption of normality of ui, equips us with the necessary tools for estimation and hypothesis testing in linear regression models, there is no loss for readers opting not to pursue the maximum likelihood method due to its slight mathematical complexity.

2.4.11. Statistical Inference in the Linear Regression Model

The realm of statistical inference in the context of the linear regression model is a fascinating one, indeed. It is a domain that is rife with complexity and nuance, yet

it is also one that offers profound insights into the nature of data and the relationships that exist within it.

The crux of statistical inference in linear regression lies in the estimation of the parameters of the model. These parameters, often denoted as beta coefficients, are the lifeblood of the model, providing the means by which the independent variables are related to the dependent variable. The estimation of these parameters is typically achieved through the method of least squares, a technique that minimizes the sum of the squared residuals, thus ensuring that the model's predictions are as close as possible to the observed values.

However, the estimation of parameters is merely the first step in the journey of statistical inference. Once the parameters have been estimated, the next step is to assess the validity of these estimates. This is where the concept of hypothesis testing comes into play. Hypothesis testing is a statistical procedure that allows us to make inferences about the population parameters based on the sample data. In the context of linear regression, hypothesis testing is often used to determine whether the estimated parameters are statistically significant, i.e., whether they are different from zero.

In addition to hypothesis testing, another critical aspect of statistical inference in linear regression is the construction of confidence intervals. Confidence intervals provide a range of values within which the true population parameter is likely to fall. They offer a measure of the uncertainty associated with the parameter estimates and are a crucial tool for understanding the precision of the estimates.
Furthermore, the assumptions underlying the linear regression model play a pivotal role in statistical inference. These assumptions, which include linearity, independence, homoscedasticity, and normality, are essential for the validity of the inference procedures. Violations of these assumptions can lead to biased or inefficient estimates, thus undermining the reliability of the model's predictions.

Hypothesis testing is a fundamental procedure in statistics that allows us to make inferences or draw conclusions about a population based on a sample of data. Here are the steps involved in hypothesis testing:

2.4.11.1. Formulate the Hypotheses

The first step in hypothesis testing is to set up the null hypothesis (H_0) and the alternative hypothesis (H_1 or H_a). The null hypothesis is a statement about the population that will be tested. The alternative hypothesis is what you might believe to be true or hope to prove true.

2.4.11.2. Choose the Significance Level (α)

The significance level, also denoted as alpha or α, is a threshold that determines when we reject the null hypothesis. Commonly used values are 0.05 (5%) and 0.01 (1%).

2.4.11.3. Select the Appropriate Test Statistic

Depending on the nature of the data and the hypothesis, select the appropriate test statistic (e.g., Z-score, t-score, F-score, etc.). The test statistics will help us decide whether to reject or fail to reject the null hypothesis.

2.4.11.4. Calculate the Test Statistic and Corresponding P-Value

The test statistics are calculated using your sample data. Once the test statistic is calculated, you can find the corresponding p-value. The p-value is the probability that you would observe a test statistic as extreme as the one calculated, assuming the null hypothesis is true.

2.4.11.5. Compare the P-Value to the Significance Level

If the p-value is less than or equal to the significance level, we reject the null hypothesis. If the p-value is greater than the significance level, we fail to reject (or retain) the null hypothesis.

2.4.11.6. Make a Decision and Interpret the Result

Based on the comparison, we decide about the hypotheses. If we rejected the null hypothesis, we could say that our sample provides enough evidence to support the alternative hypothesis. If we failed to reject the null hypothesis, we do not have enough evidence to support the alternative hypothesis. Remember, failing to reject the null hypothesis does not prove it true. It merely suggests that we do not have strong enough evidence against it. Similarly, rejecting the null hypothesis does not prove the alternative hypothesis; it suggests that the alternative may be true, and the null is unlikely.

In conclusion, statistical inference in the linear regression model is a multifaceted process that involves the estimation of parameters, hypothesis testing, the construction of confidence intervals, and the verification of model assumptions. It is a process that requires a deep understanding of statistical principles and a keen eye for detail. Yet, for those who are willing to delve into its intricacies, it offers a powerful tool for making sense of the world around us.

2.4.12. Analysis of Variance of the Linear Regression Model

Analysis of Variance (ANOVA) in a Linear Regression Model is a statistical method that helps us understand how much of the variation in our data can be explained by the model we've built. It's like a report card for our model, telling us how well it's doing.

The ANOVA table is a summary of this analysis. It breaks down the total variation in our data (Total Sum of Squares) into the part that our model can explain (Regression Sum of Squares) and the part that remains unexplained (Residual or Error Sum of Squares). Degrees of freedom, another part of the table, tell us how many values in our calculations are free to vary. It's a bit like knowing how many pieces of a puzzle we have left to place.

Table 2.1: ANOVA Table for the Two-variable Regression Model

Source of Variation	Sum of Squares	df	Mean sum of Squares
Due to regression (ESS)	$\sum \hat{y}_i^2 = \hat{\beta}_1^2 \sum X_i^2$	1	$\sum \hat{y}_i^2 = \hat{\beta}_1^2 \sum X_i^2$
Due to residuals (RSS)	$\sum \hat{u}_i^2$	$n-2$	$\frac{\sum \hat{u}_i^2}{n-2} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n-1$	

The mean squares are calculated by dividing each sum of squares by its corresponding degrees of freedom. They help us understand the average variation explained by the model and the average variation that remains unexplained. Finally, the F-statistic is a ratio of the mean square for regression to the mean square for error. It gives us a measure of how significant our model is overall. If our model is doing a good job, the F-statistic will be large, suggesting that our model is explaining a lot of the variation in the data.
In short, ANOVA in a Linear Regression Model is a tool that helps us understand how well our model is doing in explaining the variation in our data.

2.5. Self-Assessment Questions

- What is the main purpose of studying econometrics, and how does it contribute to understanding economic relationships?
- Explain the concept of the Population Regression Function (PRF) and its significance in regression analysis.
- Why is the assumption of normality of the stochastic disturbance term essential in Ordinary Least Squares (OLS) regression? How does it impact the OLS estimators' properties?
- Describe the method of OLS and its advantages as an estimator in linear regression models.
- List and explain the assumptions of the Ordinary Least Squares method. How do violations of these assumptions affect the reliability of OLS estimates?
- Discuss the properties of the Least Squares method and why it is considered a powerful estimation technique in econometrics.
- What are the measures of goodness of fit, and how do they help evaluate the performance of a regression model?
- Explain the significance of the probability distribution of the disturbance term in regression analysis. How does the choice of distribution affect the model's validity?
- Compare and contrast the Ordinary Least Squares (OLS) and Maximum Likelihood (ML) methods for estimating regression coefficients. Under what conditions do they yield identical results?
- How does statistical inference in the linear regression model enable researchers to draw meaningful conclusions from data? What role does the analysis of variance play in understanding the contributions of explanatory variables in the model?

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. 2nd Edition, McMillan.

Additional Readings

- Gujarati, D. N. Basic Econometrics, Latest edition, McGraw-Hill.
- Gujarati, D. N. Econometrics by Example. Palgrave Macmillan.
- Kennedy, P. A guide to econometrics. John Wiley & Sons.
- Rao, C. R. *Linear Statistical Inference and Its Applications*, John Wiley & Sons, New York.
- Stock, J. H., & Watson, M. W. Introduction to Econometrics. Pearson.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 03

MULTIPLE REGRESSION MODELS

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

3.1. INTRODUCTION

This unit explores the comprehensive subject of multiple regression models. Our journey will take us from simple to complex applications, beginning with models that incorporate two explanatory variables. We will delve into the heart of statistical inference within these models, exploring the various types of hypothesis testing. This includes tests of individual regression coefficients, the overall significance of our models, and the equivalency of different coefficients. Further, we will scrutinize the stability and functional form of our models, as well as specific constraints that may be placed upon them.

The interpretation of regression coefficients forms an essential part of our exploration, providing crucial insights into the relationships between our predictor and response variables. Equally important is our investigation into the relationship between partial and multiple correlation coefficients, enabling a more profound understanding of the intricate interdependencies among variables. Our focus will then shift to predictive applications of multiple regression models, exploring how they can be harnessed to predict dependent variable values based on a set of independent variables. Concluding our exploration, we will discuss the multiple coefficient of determination, a key metric indicating how effectively our regression model can predict outcomes.

Throughout this unit, our objective is to enhance your understanding of multiple regression models, empowering you to explore and interpret complex relationships within your data.

3.2. OBJECTIVES

While reading this unit, a student could aim to achieve the following objectives:

- **understand the Concept of Multiple Regression Models:** The unit begins by introducing the concept of multiple regression models using two explanatory variables. The student should aim to grasp how multiple independent variables can be incorporated into a regression model.
- **master the Techniques of Statistical Inference:** A significant portion of the chapter is dedicated to various forms of statistical inference. The student's objective should be to learn how to conduct different hypothesis tests in the context of multiple regression models, including tests on individual regression coefficients, the overall significance of the model, equality of coefficients, and more.

- **interpret Regression Coefficients:** Understanding the interpretation of regression coefficients in a multiple regression model is crucial. The student should aim to develop the skill to interpret these coefficients accurately.
- **understand Partial and Multiple Correlation:** A deep understanding of the correlation among variables is crucial in multiple regression models. The student should aim to understand the relationship between partial and multiple correlation coefficients.
- **develop Prediction Skills:** The ability to make predictions based on a multiple regression model is an essential skill in many fields. The student should aim to learn how to use the model for prediction purposes.
- **understand the Multiple Coefficient of Determination:** The student should aim to understand what the multiple coefficient of determination is, how it's calculated, and how it can be used to evaluate the effectiveness of a multiple regression model.

By achieving these objectives, a student will develop a solid understanding of multiple regression models and gain valuable skills that can be applied in various research and professional contexts.

3.3. Major Topics

- A Model with Two Explanatory Variables
- Statistical Inference in The Multiple Regression Model
- Interpretation of the Regression Coefficients
- Partial and Multiple Correlation coefficients and their Relationship
- Prediction in the Multiple Regression Model
- The Multiple Coefficient of Determination

3.4. Summary of the Units

3.4.1. A Model with Two Explanatory Variables

Building on the concept of simple regression models, a multiple regression model expands its scope to include more than one explanatory variable. For instance, consider a model with two explanatory variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \qquad (3.1)$$

This equation represents a model where Y, the dependent variable, is explained by two independent variables, X_1 and X_2, where β_0 is the Y-intercept, β_1 and β_2 are the slope coefficients associated with X_1 and X_2 respectively, and u is the error term. The coefficients β_1 and β_2 depict the average change in Y for a unit change in X_1 and X_2, respectively, holding the other variable constant. This notion of "ceteris paribus" (all other things being equal) is essential for understanding the relationship between multiple independent variables and the dependent variable.

Under the framework of the classical linear regression model (CLRM), assumptions are same as of simple linear regression model. Hence, we assume the following (detail of each assumption can be seen in section 2.4.4):

- Multiple linear regression model is a linear model (linear in the parameters.
- Fixed X values or X values are independent of the error term. Here, this means, we require:
  $$Cov\,(u_i, X_{1i}) = Cov\,(u_i, X_{2i}) = 0 \qquad (3.2)$$
- Zero mean of value of the disturbances u_i ,
  $$E(u_i | X_{1i}, X_{2i}) = 0 \quad \text{for each } i \qquad (3.3)$$
- Constant variance of u_i or homoskedastic
  $$Var\,(u_i) = \sigma^2 \qquad (3.4)$$
- There will be no autocorrelation or serial correlation, between the disturbances.
  $$Cov\,(u_i, u_j) = 0 \qquad i \neq j \qquad (3.5)$$

- The number of observations n must be greater than the number of parameters to be estimated.
- There must be variation in the values of the X variables.
- In a multiple regression model, this assumption ensures that the independent variables are not perfectly linearly related. In other words, no independent variable is a perfect linear function of other explanatory variables. If this assumption is violated, it would be impossible to separate out the individual effects of the independent variables on the dependent variable, making it impossible to estimate the individual parameters. Mathematically, this assumption can be expressed in terms of the covariance between the explanatory variables:

  $$Cov\,(X_1, X_2) = 0 \qquad (3.6)$$

  Keep in mind that we are talking only about perfect linear relationships between two or more variables. Multicollinearity does not rule out nonlinear relationships between variables.
- There is no specification bias and the model is correctly specified.

To find the OLS estimators, let us first write the sample regression function (SRF) corresponding to the PRF of equation (3.1) as follows:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i \qquad (3.7)$$

Where $\hat{u}_i$ is the residual term, the sample counterpart of the stochastic disturbance term u_i. The OLS procedure consists of choosing the values of the unknown parameters so that the residual sum of squares (RSS) $\sum \hat{u}_i^2$ is as small as possible. Symbolically,

$$\min \sum \hat{u}_i^2 = min \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2 \qquad (3.8)$$

The most straightforward procedure to obtain the estimators that will minimize equation (3.8) is to differentiate it with respect to the unknowns, set the resulting expressions to zero, and solve them simultaneously. This procedure gives the following normal equations:

$$\sum Y_i = n\,\hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} \qquad (3.9)$$

$$\sum Y_i X_{1i} = \hat{\beta}_0 \sum X_{1i} + \hat{\beta}_1 \sum X_{1i}^2 + \hat{\beta}_2 \sum X_{1i} X_{2i} \qquad (3.10)$$

$$\sum Y_i X_{2i} = \hat{\beta}_0 \sum X_{2i} + \hat{\beta}_1 \sum X_{1i} X_{2i} + \hat{\beta}_2 \sum X_{2i}^2 \qquad (3.11)$$

Solving these equations, simultaneously will give us OLS estimators of parameters β_0 , β_1 , β_2. Formulas for the OLS estimators are:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \qquad (3.12)$$

$$\hat{\beta}_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \qquad (3.13)$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \qquad (3.14)$$

About the OLS estimators, it should be noted that:

- Equations (3.13) and (3.14) are symmetrical in nature because one can be obtained from the other by interchanging the roles of $X_1$and X_2.
- The denominators of these two equations are identical
- The three-variable case is a natural extension of the two-variable case

The unbiased estimator of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3} \quad (3.15)$$

The degrees of freedom are now $(n - 3)$ because in estimating $\hat{u}_i^2$ we must first estimate β_0, β_1, and β_2, which consume 3 df.

3.4.2. Statistical Inference in the Multiple Regression Model

After surpassing the realm of the basic two-variable linear regression model, hypothesis testing takes on various intriguing forms, exemplified by the following:

3.4.2.1. Testing Hypotheses about an Individual Partial Regression Coefficient

The procedure of hypothesis testing for individual parameters in multiple linear regression model is same as of hypothesis testing for individual parameters in simple linear regression models (discussed in previous unit). Following test statistics will be used to check the significance of β_0, β_1, and β_2.

$$t = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \quad (3.16)$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \quad (3.17)$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \quad (3.18)$$

It follows the t-distribution with $n - 3$ df.

3.4.2.2. Testing the Overall Significance of the Estimated Regression Model

In this case, the null hypothesis is a joint hypothesis that β_1, and β_2 are jointly or simultaneously equal to zero. Following is the hypothesis of overall significance of the observed or estimated regression line:

$$H_0: \beta_1 = \beta_2 = 0$$

We cannot use the usual t-test to test the joint hypothesis that the true partial coefficients are zero simultaneously. However, this joint hypothesis can be tested by the analysis of variance (ANOVA) technique. Under the assumption of normal distribution of u_i, following test statistics can be employed:

$$\mathrm{F} = \frac{(\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i})/2}{\sum \hat{u}_i^2/(n-3)} = \frac{ESS/df}{RSS/df} \quad (3.19)$$

It follows F-distribution with 2 and $n - 3$ d.f. The ANOVA table in case of multiple regression model is as follow:

Table 3.1: ANOVA Table for the Three-variable Regression Model

Source of Variation	Sum of Squares	d.f.	Mean sum of Squares
Due to regression (ESS)	$\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i}$	2	$\frac{\hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i}}{2}$
Due to residuals (RSS)	$\sum \hat{u}_i^2$	$n - 3$	$\frac{\sum \hat{u}_i^2}{n-3} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n - 1$	

Generally, to test the overall significance of the regression with k-variable regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{\mathrm{k}-1} X_{\mathrm{k}-1i} + u_i \qquad (3.20)$$

Following will be the hypothesis:

H_0: $\beta_2 = \beta_3 = \beta_4 = \cdots = \beta_{\mathrm{k}-1} = 0$

H_1: not all slope coefficients are simultaneously zero

The hypothesis can be tested using the following test-statistic:

$$F = \frac{ESS/df}{RSS/df} = \frac{(R^2)/(k-1)}{(1-R^2)/(n-k)} \qquad (3.21)$$

If $F > F_{\alpha(k-1,n-k)}$, reject H_0, where $F_{\alpha(k-1,n-k)}$ is the critical F value at the α level of significance and $(k-1)$ numerator df and $(n-k)$ denominator df. Alternatively, if the P-value of the F obtained from equation (3.21) is sufficiently low, reject H_0.

3.4.2.3. Testing that Two or More Coefficients are Equal to One Another

Suppose in the multiple regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (3.20)$$

We want to test the hypothesis that the two slope coefficients β_2 and β_3 are equal:

H_0: $\beta_2 = \beta_3$ or $(\beta_2 - \beta_3) = 0$

H_0: $\beta_2 \neq \beta_3$ or $(\beta_2 - \beta_3) \neq 0$ (3.21)

Under the classical assumptions it can be shown that it follows a t-distribution. So we can use the following test-statistic:

$$t = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{se(\hat{\beta}_2 - \hat{\beta}_3)} = \frac{(\hat{\beta}_2 - \hat{\beta}_3) - (\beta_2 - \beta_3)}{\sqrt{var(\hat{\beta}_2) + var(\hat{\beta}_3) - 2cov(\hat{\beta}_2, \hat{\beta}_3)}} \qquad (3.22)$$

Now, if the computed value of t exceeds the critical t-value at the designated level of significance for given df, then we can reject H_0, otherwise, we do not reject H_0. Alternatively, if the p-value of the t-statistic is reasonably low, one can reject the null hypothesis. It should be noted that when we say that a p-value is low or reasonably low, we mean that it is less than the significance level, such as 10, 5, or 1 percent.

3.4.2.4. Testing that the Partial Regression Coefficients Satisfy Certain Restrictions

On certain occasions, economic theory might propose that the coefficients in a regression model adhere to specific linear equality restrictions. For example, let's consider the Cobb-Douglas production function.

$$Y_i = \beta_0 X_{1i}^{\beta_1} X_{2i}^{\beta_2} e^{u_i} \quad (3.23)$$

Where Y represents output, X_1 represents labor inputs and X_2 represents capital input. In the log form, equation becomes:

$$\ln Y = \alpha_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + u_i \quad (3.24)$$

Where $\alpha_0 = \ln \beta_0$, now if there are constant returns to scale (equi-proportional change in output for an equi-proportional change in the inputs), economic theory would suggest that:

$$\beta_1 + \beta_2 = 1 \quad (3.25)$$

This is an example of linear equality restriction. One approach is to check the restriction is by t-test. In this estimate equation (3.24) without taking into account the restriction given in equation (3.25), This is called the unrestricted or unconstrained regression. Following t-test can be used to test the restriction given in equation (3.25):

$$t = \frac{(\widehat{\beta}_1+\widehat{\beta}_2)-(\beta_1+\beta_2)}{se(\widehat{\beta}_2+\widehat{\beta}_3)} = \frac{(\widehat{\beta}_1+\widehat{\beta}_2)-1}{\sqrt{var(\widehat{\beta}_2)+var(\widehat{\beta}_3)+2cov(\widehat{\beta}_2,\widehat{\beta}_3)}} \quad (3.26)$$

If the t-value from equation (3.26) exceeds the critical t-value at the chosen level of significance, we reject the hypothesis of constant returns to scale, otherwise we do not reject it.

The t-test is based on the estimation of unrestricted model, whereas in another approach, test can be applied by incorporating the restriction ($\beta_1 = 1 - \beta_2$ or $\beta_2 = 1 - \beta_1$). Using any of these restriction, equation (3.24) can be written as:

$$\ln Y = \alpha_0 + (1 - \beta_2) \ln X_{1i} + \beta_2 \ln X_{2i} + u_i$$
$$\ln Y - \ln X_{1i} = \alpha_0 - \beta_2 \ln X_{1i} + \beta_2 \ln X_{2i} + u_i$$
$$\ln Y - \ln X_{1i} = \alpha_0 + \beta_2 (\ln X_{2i} - \ln X_{1i}) + u_i$$
$$\ln(Y / X_{1i}) = \alpha_0 + \beta_2 \ln(X_{2i} / X_{1i}) + u_i \quad (3.27)$$

Where Y/X_{1i} represents the output to labor ratio and X_{2i}/X_{1i} represents the capital labor ratio. To test the restriction presented in equation (3.25), we estimate the unrestricted (presented in equation 3.24) and restricted model (presented in equation 3.27). We obtain residual sum of square (RSS) from both regressions. F-test under the following test-statistic can be applied to check the restriction:

$$F = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} = \frac{(\sum \hat{u}_R^2 - \sum \hat{u}_{UR}^2)/m}{\sum \hat{u}_{UR}^2/(n-k)} = \frac{(R_{UR}^2 - R_R^2)/m}{(1-R_{UR}^2)/(n-k)} \quad (3.28)$$

Where $\sum \hat{u}_R^2$ presents the RSS of restricted model, $\sum \hat{u}_{UR}^2$ represents the RSS of unrestricted model, m represents the number of linear restrictions (1 in the present

example), R_R^2 represents the R^2 value obtained from restricted regression, and R_{UR}^2 represents the R^2 value obtained from unrestricted regression. Hence if calculated value of F-statistics is greater than critical value of F then we will null hypothesis of constant returns to scale.

3.4.2.5. Testing the Stability of the Regression Model over time or in Different Cross-Sectional Units

In the regression with time series data, there may be a structural change in the relationship between the regressand Y and the regressors. By structural change, we mean that the values of the parameters of the model do not remain the same through the entire time period. Now the possible differences, that is, structural changes, may be caused by differences in the intercept or the slope coefficient or both. Chow test developed by Chow (1960) can be used to check test structural break. Suppose the data is divided in to groups with n_1 observations in the first group and n_2 observations in the second group. So, we can have three possible regressions:

$$Y_t = \lambda_0 + \lambda_1 X_t + u_{1t} \quad \text{for } n_1 \text{ observations} \quad (3.29)$$

$$Y_t = \gamma_0 + \gamma_1 X_t + u_{2t} \quad \text{for } n_2 \text{ observations} \quad (3.30)$$

$$Y_t = \alpha_0 + \alpha_1 X_t + u_t \quad \text{for } n_1 + n_2 \text{ observations} \quad (3.31)$$

Chow test assumes that:

- $u_{1t} \sim N(0, \sigma^2)$ and $u_{2t} \sim N(0, \sigma^2)$. That is, the error term in the subperiod regressions are normally distributed with the same (homoscedastic) variance σ^2.
- The two error terms u_{1t} and u_{2t} are independently distributed.

The mechanics of Chow-test are as follows:

- Estimate regression (3.31) which if there is no parametric instability, and obtain RSS_3 with df $(n_1 + n_2 - k)$, where k is the number of parameters estimated. We call RSS_3 the restricted residual sum of squares (RSS_R) because it is obtained by imposing restriction $\lambda_0 = \gamma_0$ and $\lambda_1 = \gamma_1$.
- Estimate regression (3.29) and obtain its residual sum of square, RSS_1 with df $(n_1 - k)$.
- Estimate regression (3.30) and obtain its residual sum of square, RSS_2 with df $(n_2 - k)$.
- Since the two sets of samples are deemed independent, we can add RSS_1 and RSS_2 to obtain what may be called the unrestricted residual sum of squares ($RSS_{\text{UR}} = RSS_1 + RSS_2$), with df $(n_1 + n_2 - 2k)$.
- Now the idea behind the Chow test is that if in fact there is no structural change (i.e., regressions [3.29] and [3.30] are essentially the same), then the

RSS_{R} and RSS_{UR} should not be statistically different. Therefore, if we form the following ratio:

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n_1+n_2-2k)} \qquad (3.32)$$

then Chow has shown that under the null hypothesis the regressions (3.29) and (3.30) are (statistically) the same (i.e., no structural change or break) and the F ratio given above follows the F distribution with k and $(n_1 + n_2 - 2k)$ df in the numerator and denominator, respectively.

- If the computed F value does not exceed the critical F value at chosen level of significance then we do not reject the null hypothesis of parameter stability (i.e., no structural change). In this case we may be justified in using the pooled regression (3.31).

There are some caveats about the Chow test that must be kept in mind:

- To conduct the test successfully, it is crucial to ensure that the underlying assumptions are met. For instance, it is necessary to verify whether the error variances in regressions (3.29) and (3.30) are equal.
- The Chow test will only indicate whether there is a difference between the two regressions (3.29) and (3.30), but it will not specify whether the difference is due to the intercepts, the slopes, or both.
- The Chow test relies on the assumption that we have knowledge of the point(s) of structural break. Nevertheless, if determining the actual occurrence of the structural change becomes challenging, alternative methods may need to be employed.

3.4.2.6. Testing the Functional Form of the Regression Model

The decision of whether to opt for a linear regression model or a log-linear regression model remains an enduring question in empirical analysis. To address this, we can employ a test proposed by MacKinnon, White, and Davidson (1983), commonly referred to as the MWD test, which helps us make a choice between the two models. A similar test is proposed in Bera and Jarque (1982). For the test, following is the null and alternative hypothesis:

H_0: Linear Model: Y is a linear function of regressors, the X's

H_1: Log-Linear Model: $\ln Y$ is a linear function of logs of regressors, the logs of X's

The MWD test involve the following steps:

- Estimate the linear model and obtain the estimated values of Y; Call them Yf (i.e. $\hat{Y}$).
- Estimate the log-linear model and obtain the estimated values of $\ln Y$; call them $\ln f$ (i.e. $\widehat{\ln Y}$).
- Obtain $Z_1 = (\ln Yf - \ln f) = (\ln \hat{Y} - \widehat{\ln Y})$.

- Regress Y on X's and Z_1 obtained in previous step. Reject H_0 if the coefficient of Z_1 is statistically significant by the usual t-test.
- Obtain $Z_2 = (antilog\ of\ \ln Yf - \ln f) = \left(antilog\ of\ \ln \hat{Y} - \ln \widehat{Y}\right)$.
- Regress log of Y on the log's of X's and Z_2. Reject H_1 if the coefficient of Z_2 is statistically significant by the usual t-test.

3.4.3. Interpretation of the Regression Coefficients

To comprehend the model's interpretation, we initially examined how child mortality (CM) behaves concerning per capita GNP ($PGNP$), observing a negative impact of $PGNP$ on CM, as anticipated. Now, let's introduce female literacy, represented by the female literacy rate (FLR). We also expect FLR to have a negative influence on CM. When incorporating both variables into our model, we must isolate and estimate the individual (partial) regression coefficients of each regressor to understand their respective effects.

$$CM_i = \beta_0 + \beta_1 PGNP_i + \beta_2 FLR_i + u_i \qquad (3.32)$$

By using the data of different countries, consider following is the estimated model:

$$\widehat{CM}_i = 263.6416 - 0.0056 PGNP_i - 2.2316 FLR_i \qquad (3.33)$$
$$se = (11.5932)\quad (0.0019)\qquad (0.2099)$$
$$R^2 = 0.7077 \qquad \bar{R}^2 = 0.6981$$

Let's now interpret these regression coefficients. The value of -0.0056 represents the partial regression coefficient of PGNP, indicating that while holding the influence of FLR constant, an increase of one dollar in per capita GNP, on average, results in a decrease of 0.0056 units in child mortality. To provide a more economically meaningful interpretation, if per capita GNP rises by a thousand dollars, on average, the number of deaths of children under the age of 5 decreases by approximately 5.6 per thousand live births.

The coefficient of -2.2316 reveals that while keeping the influence of PGNP constant, an increase of one percentage point in the female literacy rate leads to an average reduction of about 2.23 deaths per thousand live births for children under the age of 5. The intercept value, approximately 263, can be mechanically interpreted as the child mortality rate when both PGNP and FLR are set to zero. However, it's essential to approach this interpretation cautiously. It suggests that if both PGNP and FLR were at zero, the child mortality rate would be around 263 deaths per thousand live births. Naturally, such an interpretation should be taken with caution, as it simply implies that if both regressors were absent, the child mortality rate would be relatively high, which aligns with practical expectations.
The R-squared value of approximately 0.71 indicates that about 71 percent of the variation in child mortality can be explained by PGNP and FLR, which is relatively high considering that the maximum possible value for R-squared is 1. In summary,

the regression results appear to be sensible and provide valuable insights into the relationship between the variables.

Before moving forward, let's consider the scenario where we want to determine the effect on the child mortality rate when both PGNP and FLR are increased simultaneously. Suppose per capita GNP increases by one dollar, and at the same time, the female literacy rate goes up by one percentage point. To ascertain the impact of this simultaneous change on the child mortality rate, we simply need to multiply the coefficients of PGNP and FLR by their respective proposed changes and add the resulting terms. In our specific example, this calculation yields the following result:

$$-0.0056(1) - 2.2316(1) = -2.2372$$

That is, as a result of this simultaneous change in PGNP and FLR, the number of deaths of children under age 5 would go down by about 2.24 deaths.

Now if the regression is carried out with the standardized variables (a variable is said to standardized if it is expressed in terms of deviation from its mean and divided by its standard deviation). Consider following is the estimated model with standardized variables:

$$\widehat{CM}_i^* = -0.2026PGNP_i^* - 0.7639FLR_i^* \qquad (3.34)$$
$$se = (0.0713) \quad (0.0713)$$
$$R^2 = 0.7077$$

Not that variables with * represent standardized variables and in model with standardized variables, there is no intercept. From this regression analysis, you can observe that while keeping FLR constant, a one-standard-deviation increase in PGNP results, on average, in a 0.2026 standard deviation decrease in CM. Likewise, holding PGNP constant, a one-standard-deviation increase in FLR leads, on average, to a 0.7639 standard deviation decrease in CM. In relative terms, female literacy has a more significant impact on child mortality than per capita GNP. This highlights the advantage of using standardized variables since standardization equalizes all variables by giving them zero means and unit variances, allowing for a fair comparison of their respective impacts.

3.4.4. Partial and Multiple Correlation Coefficients and their Relationship

Coefficient of correlation r as a measure of the degree of linear association between two variables. For the three-variable regression model we can compute three correlation coefficients: r_{12} (correlation between Y and X_2), r_{13} (correlation coefficient between Y and X_3), and r_{23} (correlation coefficient between X_2 and X_3); notice that we are letting the subscript 1 represent Y for notational convenience. These correlation coefficients are called gross or simple correlation coefficients, or

correlation coefficients of zero order. These coefficients can be computed using the formula:

$$r_{12} = \frac{n\sum X_{2i}Y_i - (\sum X_{2i})(\sum Y_i)}{\sqrt{[(n\sum X_{2i}^2 - (\sum X_{2i})^2)][(n\sum Y_i^2 - (\sum Y_i)^2)]}}$$
$$r_{13} = \frac{n\sum X_{3i}Y_i - (\sum X_{3i})(\sum Y_i)}{\sqrt{[(n\sum X_{3i}^2 - (\sum X_{3i})^2)][(n\sum Y_i^2 - (\sum Y_i)^2)]}}$$
$$r_{23} = \frac{n\sum X_{2i}X_{3i} - (\sum X_{2i})(\sum X_{3i})}{\sqrt{[(n\sum X_{2i}^2 - (\sum X_{2i})^2)][(n\sum X_{3i}^2 - (\sum X_{3i})^2)]}} \quad (3.35)$$

Simple correlation lies between the limits of −1 and +1; that is, $-1 \leq r \leq 1$. In general, r_{12} is not likely to reflect the true degree of association between Y and X_2 in the presence of X_3. As a matter of fact, it is likely to give a false impression of the nature of association between Y and X_2. Therefore, what we need is a correlation coefficient that is independent of the influence, if any, of X_3 on X_2 and Y. Such a correlation coefficient can be obtained and is known appropriately as the partial correlation coefficient. Conceptually, it is similar to the partial regression coefficient. We define $r_{12.3}$ as partial correlation coefficient between Y and X_2, holding X_3 constant, $r_{13.2}$ as partial correlation coefficient between Y and X_3, holding X_2 constant, $r_{23.1}$ as partial correlation coefficient between X_2 and X_3, holding X_2 constant. These partial correlations can be easily obtained from the simple or zero-order, correlation coefficients as follows:

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$
$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$
$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}} \quad (3.36)$$

The above partial correlations are called first-order correlation coefficients. In the two variable case r measures the degree of (linear) association between the dependent variable Y and the single explanatory variable X. Beyond the two-variable case, we observe the following:

- Even if $r_{12} = 0$, $r_{12.3}$ will not be zero unless r_{13} or r_{23} or both are zero.
- If $r_{12} = 0$, and r_{13} and r_{23} are nonzero and are of the same sign, $r_{12.3}$ will be negative, whereas if they are of the opposite signs, it will be positive.
- The terms $r_{12.3}$ and r_{12} (and similar comparisons) need not have the same sign.
- In the two-variable case we have seen that r^2 lies between 0 and 1. Similarly, for partial correlations we can write:

$$0 \leq r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \leq 1$$

- The fact that Y and X_3 and X_2 and X_3 are uncorrelated ($r_{12} = r_{23} = 0$) does not mean that Y and X_2 are uncorrelated.

The term $r^2_{12.3}$ may be called the coefficient of partial determination and may be interpreted as the proportion of the variation in Y not explained by the variable X_3 that has been explained by the inclusion of X_2 into the model. Further, the relationships between R^2 with simple correlation coefficients and partial correlation coefficients can be expressed as:

$$R^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$
$$R^2 = r_{12}^2 + (1 - r_{12}^2)r_{13.2}^2$$
$$R^2 = r_{13}^2 + (1 - r_{13}^2)r_{12.3}^2 \quad (3.37)$$

In concluding this section, consider the following: It was stated previously that R^2 will not decrease if an additional explanatory variable is introduced into the model, which can be seen clearly from second equation of (3.37). This equation states that the proportion of the variation in Y explained by X_2 and X_3 jointly is the sum of two parts: the part explained by X_2 alone ($= r_{12}^2$) and the part not explained by X_2 ($= 1 - r_{12}^2$) times the proportion that is explained by X_3 after holding the influence of X_2 constant ($= r_{13.2}^2$). Now $R^2 > r_{12}^2$ so long as $r_{13.2}^2 > 0$. At worst, $r_{13.2}^2$ will be zero, in which case $R^2 = r_{12}^2$.

3.4.5. Prediction in the Multiple Regression Model

Similar to estimated two-variable regression model, estimated multiple regression model too can be used for (1) mean prediction, that is, predicting the point on the population regression function (PRF), as well as for (2) individual prediction, that is, predicting an individual value of Y given the value of the regressors.

3.4.6. The Multiple Coefficient of Determination

The coefficient of determination is:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum u_i^2}{\sum y_i^2} \quad (3.38)$$

Now $\sum y_i^2$ is independent of the number of X variables in the model because it is simply $(Y_i - \bar{Y})^2$. The RSS, $\sum u_i^2$, however, depends on the number of regressors present in the model. Intuitively, it is clear that as the number of X variables increases, $\sum u_i^2$ is likely to decrease (at least it will not increase); hence R^2 as defined in Eq. (3.35) will increase. In view of this, in comparing two regression models with the same dependent variable but differing number of X variables, one should be very wary of choosing the model with the highest R^2. Adjusted R^2 (denoted by $\bar{R}^2$) is often used to address this issue.

$$\bar{R}^2 = 1 - \left(\frac{\frac{\sum u_i^2}{n-k}}{\frac{\sum y_i^2}{n-1}}\right) = 1 - \left[(1-R^2)\left(\frac{n-1}{n-k}\right)\right] = 1 - \frac{\hat{\sigma}^2}{S_Y^2} \quad (3.39)$$

It is immediately apparent from equation (3.36) that:

- for $k > 1$, $\bar{R}^2 < R^2$ which implies that as the number of X variables increases, the adjusted R^2 increases less than the unadjusted R^2.
- $\bar{R}^2$ can be negative, R^2 is necessarily non-negative. That is if $R^2 = 1$, then $\bar{R}^2 = R^2 = 1$. When $R^2 = 0$, then $\bar{R}^2 = (1-k)/(n-k)$ can be negative if $k > 1$. 0 In case $\bar{R}^2$ turns out to be negative in an application, its value is taken as zero.

According to Theil (1978),

> *...it is good practice to use $\bar{R}^2$ rather than R^2 because R^2 tends to give an overly optimistic picture of the fit of the regression, particularly when the number of explanatory variables is not very small compared with the number of observations.*

Further, it is crucial to note that in comparing two models on the basis of the coefficient of determination, whether adjusted or not, the sample size n and the dependent variable must be the same

In concluding this section, a word of caution is necessary: Some researchers may fall into the trap of maximizing $\bar{R}^2$, meaning they choose the model that yields the highest $\bar{R}^2$ value. However, this approach can be perilous because our primary objective in regression analysis is not solely to achieve a high $\bar{R}^2$, but rather to obtain reliable estimates of the true population regression coefficients and draw meaningful statistical inferences about them. In empirical analysis, it is not uncommon to achieve a very high $\bar{R}^2$ but discover that some of the regression coefficients are either statistically insignificant or have signs that contradict a priori expectations. Hence, researchers should prioritize the logical or theoretical relevance of the explanatory variables to the dependent variable and their statistical significance. If, in this process, a high $\bar{R}^2$ is obtained, it is certainly beneficial. Conversely, if $\bar{R}^2$ is low, it does not necessarily imply that the model is inadequate (Achen, 1982; Granger & Newbold, 1976). According to Goldberger (1991):

> *From our perspective, R^2 has a very modest role in regression analysis, being a measure of the goodness of fit of a sample LS [least-squares] linear regression in a body of data. Nothing in the CR [CLRM] model requires that R^2 be high. Hence a high $R^2$is not evidence in favor of the model and a low R^2 is not evidence against it.*

3.5. Self-Assessment Questions

- Can you explain the concept of a multiple regression model and its components? Why is it advantageous to use two or more explanatory variables?
- Can you describe the process of testing hypotheses about an individual partial regression coefficient in a multiple regression model? How about testing the overall significance of the estimated regression model?
- How do you interpret the regression coefficients in a multiple regression model? What does a positive or negative coefficient imply?
- Can you distinguish between partial and multiple correlation coefficients? How are they related, and what does each represent in a multiple regression model?
- How do you use a multiple regression model to predict the value of a dependent variable? What steps would you follow to perform this prediction?
- Can you define the multiple coefficients of determination and explain its significance? How can it be used to assess the quality of a multiple regression model?
- Can you provide an example from a real-world scenario where you could apply a multiple regression model to solve a problem or answer a question? How would you interpret the results?
- How would you test that two or more coefficients are equal in a multiple regression model?
- Using two subsets of data from different time periods or cross-sectional units, can you test the stability of the regression model over time or across units?

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Achen, C. H. Interpreting and Using Regression, Sage Publications, Beverly Hills, Calif., 58–67.
- Bera, A. K., & Jarque, C. M. Model Specification Tests: A Simultaneous Approach. Journal of Econometrics, 20, 59–82.
- Chow, Gregory C. (1960). Tests of equality between sets of coefficients in two linear regressions. Econometrica, 28(3), 591–605.
- Goldberger, A. S. A Course in Econometrics, Harvard University Press, Cambridge, Mass., .
- Granger, C. & Newbold, P. (1976). R2 and the Transformation of Regression Variables. Journal of Econometrics, 4, 205–210.
- Gujarati, D. N. Basic Econometrics, Latest edition, McGraw-Hill.
- Gujarati, D. N. Econometrics by Example. Palgrave Macmillan.
- MacKinnon, J., White, H., & Davidson, R. (1983). Tests for Model Specification in the Presence of Alternative Hypothesis; Some Further Results. Journal of Econometrics, 21, 53–70.
- Stock, J. H., & Watson, M. W. Introduction to Econometrics. Pearson.
- Theil, H. Introduction to Econometrics, Prentice Hall, Englewood Cliffs, NJ, p. 135.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 04

THE MATRIX APPROACH TO LINEAR REGRESSION MODEL

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

4.1. INTRODUCTION

Welcome to Unit 4, where we delve into the intricacies of the k-Variable Linear Regression Model. Throughout this Unit, we will explore various fundamental aspects of linear regression using matrix notations. Starting with an understanding of the assumptions that underpin this model, we will then proceed to grasp the core concepts of OLS Estimation and the properties of OLS estimators. Hypothesis testing and the analysis of variance will also be presented in the context of matrix notation. Additionally, we will explore the significance and applications of the correlation matrix. So, let's embark on this journey to enhance our comprehension of the k-Variable Linear Regression Model and its key components.

4.2. Objectives

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

- **understand the k-Variable Linear Regression Model:** Gain a comprehensive understanding of the k-variable linear regression model, its components, and how it is represented using matrix notations.
- **grasp Assumptions in Matrix Notations:** Familiarize yourself with the assumptions underlying the linear regression model when expressed in matrix form, which form the basis for reliable estimation.
- **master OLS Estimation and Properties:** Learn the Ordinary Least Squares (OLS) estimation technique, and explore the essential properties of OLS estimators, enabling accurate parameter estimation in regression models.
- **perform Hypothesis Testing in Matrix Notations:** Develop the skills to conduct hypothesis testing in the context of matrix notations, providing insights into the significance of various model parameters.
- **Analyze Variance using Matrix Notation:** Discover how to analyze variance in the linear regression model when represented in matrix notation, gaining insights into the distribution of the error terms.
- **explore the Correlation Matrix:** Explore the significance and applications of the correlation matrix, understanding its role in measuring the relationships between variables.
- **apply Matrix Notations in Linear Regression:** Learn to apply matrix notations effectively in various scenarios, enhancing your ability to work with complex regression models and large datasets.

- **enhance Analytical and Interpretive Skills:** Develop strong analytical skills to interpret regression results accurately, enabling you to draw meaningful insights from empirical data.

- **build a Solid Foundation in Regression Analysis:** Acquire a solid foundation in regression analysis, providing a valuable skillset for conducting empirical research and data-driven decision-making in various fields.

- **prepare for Advanced Topics:** Lay the groundwork for advanced topics in econometrics, statistics, and data science, setting the stage for further academic and professional growth.

By mastering these objectives, you will be well-equipped to navigate the intricacies of the k-Variable Linear Regression Model and apply matrix notations effectively in practical data analysis and research endeavors.

4.3. Major Topics

- The k-Variable Linear Regression Model
- Assumptions of Linear Regression model in Matrix Notations
- OLS Estimation and Properties of OLS Estimators
- Hypothesis Testing in Matrix Notations
- Analysis of Variance in Matrix Notation
- The Correlation Matrix

4.4. Summary of the Units

4.4.1. The k-Variable Linear Regression Model

Now we generalize the two- and three-variable linear regression model, with k-variable Population Regression Function (PRF) involving the dependent variable Y and $k-1$ explanatory variables $X_2, X_3, \dots, X_k$. The model can be written as:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \qquad (4.1)$$

Where, β_1 is intercept, β_2 to β_k are partial slope coefficients, u is the stochastic disturbance term and i represents ith observation, n being size of the population. The model presented in (4.1) give us mean value of Y conditional upon the fixed values of explanatory variables $[E(Y_i \mid X_{2i}, X_{3i}, \dots, X_{ki})]$.

Expressed as Equation (4.1), we compactly represent the following set of n simultaneous equations:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\ &\vdots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n \end{aligned} \qquad (4.2)$$

Now, we shall present an alternative, yet more enlightening form of the system of equations (4.2) as follows:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$\boldsymbol{Y} \quad = \quad \boldsymbol{X} \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{U} \qquad (4.3)$$

Where $\boldsymbol{Y}$ is $n \times 1$ column vector of observations on the dependent variable Y, $\boldsymbol{X}$ is $n \times k$ matrix giving n observations on $k-1$ variables X_2 to X_k (this is also known as data matrix), $\boldsymbol{\beta}$ is $k \times 1$ column vector of unknown parameters $\beta_2, \beta_3, \dots, \beta_k$, and $\boldsymbol{U}$ is $n \times 1$ column vector of n disturbances u_i. The matrix representation of the general (k-variable) linear regression model is denoted as System (4.3). In simple form it can be written as:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U} \quad (4.4)$$

For the purpose of estimation, we may use the method of least squares (OLS) of the method of maximum likelihood (ML). But as noted in unit 3, these two methods yield identical estimates of the regression coefficients.

4.4.2. Assumptions of Linear Regression Model in Matrix Notations

The assumptions underlying the classical linear regression model (CLRM) are as follow:

- **Assumption 1:** The expected value of disturbance vector $\boldsymbol{U}$, that is, of each of its element, is zero $E(\boldsymbol{U}) = 0$ or $E(u_i) = 0$ for each $i = 1,2,3, \dots, n$.

$$E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (4.5)$$

- **Assumption 2:** $E(\boldsymbol{UU'}) = \sigma^2\boldsymbol{I}$ represent two assumptions, $E(u_i u_j) = 0$ for $i \neq j$ indicating no serial correlation and $E(u_i u_j) = \sigma^2$ for $i = j$ indicating no heteroscedasticity (having homoscedasticity).

$$E(\boldsymbol{UU'}) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}[u_1 \quad u_2 \quad u_3 \quad u_4] = E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_k^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_k^2) \end{bmatrix} \quad (4.6)$$

According to the assumption of homoscedasticity $E(u_i u_j) = \sigma^2$ for $i = j$ and the assumption of no autocorrelation $E(u_i u_j) = 0$ for $i \neq j$. We can write the above matrix as follow:

$$E(\boldsymbol{UU'}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2\boldsymbol{I} \quad (4.7)$$

Where $\boldsymbol{I}$ is an $n \times n$ identity matrix. Matrix presented in (4.6) is called the variance-covariance of the disturbances u_i. Elements along the diagonal present variances, whereas elements off the diagonal present covariances. Further, it should be noted that variance-covariance matrix is symmetric.

- **Assumption 3:** The matrix $\boldsymbol{X}$ is non-stochastic, that is, it consists of fixed numbers.

- **Assumption 4:** The matrix $\boldsymbol{X}$ has full column rank equal to $\rho(\boldsymbol{X}) = k$. This means that the columns of the $\boldsymbol{X}$ matrix are linearly independent; that is, there is no exact linear relationship among the $\boldsymbol{X}$ variables. In other words there is no multicollinearity.
- **Assumption 5:** The vector $\boldsymbol{U}$ has multivariate normal distribution i.e., $\boldsymbol{U} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ or $u_i \sim N(0, \sigma^2)$. The assumption is very important for hypothesis testing.

4.4.3. OLS Estimation and Properties of OLS Estimators

Before deriving the Ordinary Least Squares (OLS) estimate of $\boldsymbol{\beta}$, let's begin by expressing the k-variable sample regression function (SRF):

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + \hat{u}_i \qquad (4.8)$$

The above equation can be written in the matrix form as follow:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\ \widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{U}}$$

$$\boldsymbol{Y} = \boldsymbol{X}\,\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{U}} \qquad (4.9)$$

Where $\widehat{\boldsymbol{\beta}}$ is a k-element column vector of OLS estimators of the regression coefficients. Similar to two- and three-variable models, OLS estimators can be obtained my minimizing:

$$\sum u_i^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_2 - \cdots - \hat{\beta}_k X_k\right)^2 \qquad (4.10)$$

Where, $\sum u_i^2$ represents the residual sum of square (RSS). In matrix notation $\sum u_i^2$ can be obtained by $\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}$.

$$\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}} = [\hat{u}_1 \quad \hat{u}_2 \quad \ldots \quad \hat{u}_n] \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_n \end{bmatrix} = u_1^2 + u_2^2 + \cdots + u_n^2 = \sum u_i^2 \qquad (4.11)$$

Now from equation (4.9)

$$\widehat{\boldsymbol{U}} = \boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{\beta}} \qquad (4.12)$$

Therefore, equation (4.11) can be written as:

$$\begin{aligned} \widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}} &= (\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{\beta}})'\left(\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}\right) \\ &= (\boldsymbol{Y}' - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}')\left(\boldsymbol{Y} - \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}\right) && \therefore \left(\boldsymbol{X}\,\widehat{\boldsymbol{\beta}}\right)' = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}' \\ &= (\boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} && \therefore\ \boldsymbol{Y}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{2}\,\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}} && (4.13) \end{aligned}$$

Equation (4.13) is the matrix representation of equation (4.10). The OLS estimators can be obtained by differentiating equation (4.13) with respect to $\widehat{\boldsymbol{\beta}}$.

$$\frac{\partial(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}})}{\partial\widehat{\boldsymbol{\beta}}} = \frac{\partial(\boldsymbol{Y}'\boldsymbol{Y}-2\,\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}+\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}})}{\partial\widehat{\boldsymbol{\beta}}}$$

$$= \frac{\partial(\boldsymbol{Y}'\boldsymbol{Y})}{\partial\widehat{\boldsymbol{\beta}}} - \frac{\partial(-2\,\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y})}{\partial\widehat{\boldsymbol{\beta}}} + \frac{\partial(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\,\widehat{\boldsymbol{\beta}})}{\partial\widehat{\boldsymbol{\beta}}}$$

$$= \mathbf{0} - 2\,(\boldsymbol{X}'\boldsymbol{Y}) + 2(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}}$$

Therefore, we can write:

$$2(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = 2(\boldsymbol{X}'\boldsymbol{Y})$$

$$(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{Y})$$

Pre-multiplying both sides by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$

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

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}) \tag{4.14}$$

Equation (4.14) is a fundamental result of the OLS theory in matrix notation. Further, using the matrix approach it is easy to write the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$.

$$\text{var-cov}\,(\widehat{\boldsymbol{\beta}}) = E\big[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\big]\big[\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\big]' \tag{4.15}$$

The variance-covariance matrix can be written as:

$$\text{var-cov}\,(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} var(\hat{\beta}_1) & cov(\hat{\beta}_1,\hat{\beta}_2) & \cdots & cov(\hat{\beta}_1,\hat{\beta}_k) \\ cov(\hat{\beta}_2,\hat{\beta}_1) & var(\hat{\beta}_2) & \cdots & cov(\hat{\beta}_2,\hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k,\hat{\beta}_1) & cov(\hat{\beta}_k,\hat{\beta}_2) & \cdots & var(\hat{\beta}_k) \end{bmatrix}$$

$$\text{var-cov}\,(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{4.16}$$

Similarly, for k-variable regression model, the unbiased estimator of σ^2 can be written as:

$$\hat{\sigma}^2 = \frac{\sum\hat{u}^2}{n-k} = \frac{\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}}{n-k} \tag{4.17}$$

The BLUE property of OLS estimators can be generalized for the vector of estimators obtain through the equation (4.14). $\widehat{\boldsymbol{\beta}}$ is linear (each of its elements is a linear function of Y, the dependent variable). $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, that is, the expected value of each element of $\widehat{\boldsymbol{\beta}}$ is equal to the corresponding element of the true $\boldsymbol{\beta}$, and in the class of all linear unbiased estimators of $\boldsymbol{\beta}$, the OLS estimator $\widehat{\boldsymbol{\beta}}$ has minimum variance.

4.4.4. Hypothesis Testing in Matrix Notations

For hypothesis testing we assume that each u_i follows the normal distribution with zero mean and constant variance σ^2. In matrix notation, we can write:

$$\boldsymbol{U} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}) \tag{4.18}$$

Where, $\mathbf{0}$ is the null vector of dimension $n \times 1$. Likewise, each element of $\widehat{\boldsymbol{\beta}}$ is normally distributed with mean equal to the corresponding element of true $\boldsymbol{\beta}$ and variance given by σ^2 times the appropriate diagonal element of the inverse matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

$$\widehat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}] \quad (4.19)$$

Since in practice σ^2 is unknown, it is estimated by $\hat{\sigma}^2$ by using equation (4.17). Each element of $\widehat{\boldsymbol{\beta}}$ follows the t-distribution with $n-k$ df. Thus t-distribution therefore be used to test hypotheses about the true $\boldsymbol{\beta}$. The test statistic is as follow:

$$t = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)} \quad (4.20)$$

4.4.5. Analysis of Variance in Matrix Notation

The ANOVA technique can be easily extended to the k-variable case. TSS, RSS and ESS in the matrix form can be obtained by using the following formulas:

$$TSS: \quad \sum y_i^2 = \boldsymbol{Y}'\boldsymbol{Y} - n\bar{Y}^2$$
$$ESS: \quad \hat{\beta}_2 \sum y_i x_{2i} + \cdots + \hat{\beta}_k \sum y_i x_{ki} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2$$
$$RSS: \quad \widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}} = \boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} \quad (4.21)$$

The degree of freedoms with these sums are $n-1$, $k-1$, and $n-k$, respectively. Given the values of TSS, ESS, and RSS one can calculate the value of R^2 using the following formula:

$$R^2 = \frac{ESS}{TSS} = \frac{\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2}{\boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}} \quad (4.22)$$

Assuming that the disturbances u_i are normally distributed and the null hypothesis is $\beta_2 = \beta_3 = \cdots = \beta_k = 0$. The test statistics for the hypothesis is:

$$F = \frac{(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2)/(\boldsymbol{k-1})}{(\boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y})/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \quad (4.23)$$

The ANOVA tables for the above test statistics can be written as:

Table 4.1: Matrix formulation of the ANOVA Table for k-variable linear regression model

Source of Variation	Sum of Squares	d.f.	Mean sum of Squares
Due to regression (ESS) [Due to $X_1, X_2, \ldots, X_k$]	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2$	$k-1$	$\frac{\widehat{\boldsymbol{\beta}}\boldsymbol{X}'\boldsymbol{Y} - n\bar{Y}^2}{k-1}$
Due to residuals (RSS)	$\boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}$	$n-k$	$\frac{\boldsymbol{Y}'\boldsymbol{Y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{Y}}{n-k}$
Total (TSS)	$\boldsymbol{Y}'\boldsymbol{Y} - n\bar{Y}^2$	$n-1$	

In terms of R^2, the ANOVA table can be presented as:

Table 4.2: Matrix formulation of the ANOVA Table in matrix form in terms of R^2

Source of Variation	Sum of Squares	d.f.	Mean sum of Squares
Due to regression (ESS) [Due to $X_1, X_2, \dots, X_k$]	$R^2(\boldsymbol{Y'Y} - n\bar{Y}^2)$	$k - 1$	$\frac{R^2(\boldsymbol{Y'Y} - n\bar{Y}^2)}{k-1}$
Due to residuals (RSS)	$(1 - R^2)(\boldsymbol{Y'Y} - n\bar{Y}^2)$	$n - k$	$\frac{(1-R^2)(\boldsymbol{Y'Y} - n\bar{Y}^2)}{n-k}$
Total (TSS)	$\boldsymbol{Y'Y} - n\bar{Y}^2$	$n - 1$	

4.4.6. The Correlation Matrix

In the preceding unit, we encountered the zero-order, or simple, correlation coefficients r_{12}, r_{13}, r_{23} along with the partial, or first-order, correlations $r_{12.3}, r_{13.2}$, $r_{23.1}$, and their interrelationships. For the k-variable scenario, there will be a total of $k(k-1)/2$ zero-order correlation coefficients. These $k(k-1)/2$ correlations can be presented in the correlation matrix $\boldsymbol{R}$ as follows:

$$\boldsymbol{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & r_{22} & r_{23} & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & \cdots & r_{kk} \end{bmatrix} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & \cdots & 1 \end{bmatrix} \qquad (4.24)$$

Where the subscript 1, as before, denotes the dependent variable Y. Further, by using the formulas presented in previous unit, one can obtain the correlation coefficients of first order or higher order.

4.5. Self-Assessment Questions

- Define the k-Variable Linear Regression Model and explain its significance in statistical analysis.
- What are the key assumptions of the linear regression model when represented in matrix notations? How do these assumptions affect the reliability of the regression results?
- Describe the Ordinary Least Squares (OLS) estimation method and its primary purpose in regression analysis.
- What are the essential properties of OLS estimators, and how do they contribute to the accuracy of parameter estimation?
- How can hypothesis testing be conducted in the context of matrix notations, and what insights can be gained from such tests in linear regression?
- Explain the concept of analysis of variance in matrix notation and discuss its role in understanding the distribution of error terms in regression.
- What is the correlation matrix, and how is it calculated in the context of linear regression? Describe its applications and importance in interpreting relationships between variables.
- Demonstrate how matrix notations can be effectively applied in linear regression scenarios, using specific examples.
- Analyze a given dataset with multiple variables and perform k-Variable Linear Regression using matrix notations to estimate model parameters.
- Discuss the practical implications and limitations of the k-Variable Linear Regression Model in real-world data analysis.
- How can you leverage the knowledge gained from this unit to conduct empirical research or make data-driven decisions in your chosen field?
- Compare and contrast the k-Variable Linear Regression Model with other regression techniques, highlighting their strengths and weaknesses.
- Propose a research question in your area of interest where the k-Variable Linear Regression Model could be applied and outline the steps you would take to analyze the data using matrix notations.
- Reflect on the key concepts and skills you have learned in this chapter and identify areas where you may need further practice or study.
- Consider the broader implications of understanding regression analysis in matrix notations and how it can contribute to your academic and professional development.

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Griffiths, W. E., Hill, R. C., & Judge, G. G. *Learning and Practicing Econometrics*, Wiley.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 05

MULTICOLLINEARITY

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

5.1. INTRODUCTION

This unit explores the intricate topic of multicollinearity, a common issue in multiple regression analysis where predictor variables are closely correlated with one another. We'll begin with an exploration of the nature of multicollinearity, delving into its origins and significance in statistical analysis. Moving forward, we'll discuss how to estimate variables in scenarios where multicollinearity is present and examine the implications of such correlation on the precision of estimates and the reliability of statistical inference. This is essential to understand as multicollinearity can often distort the individual effect of predictor variables and lead to wider confidence intervals.

We'll then dive into the methods for detecting multicollinearity, focusing on key identifiers and statistical measures such as variance inflation factors (VIF) and correlation matrices. Finally, the chapter concludes with a review of various remedial measures available to tackle multicollinearity. From data transformations to variable selection, we'll discuss the pros and cons of these techniques and provide insights on how best to manage and mitigate the effects of multicollinearity in your regression models.

5.2. OBJECTIVES

By the end of this Unit, students should be able to:

- **understanding Multicollinearity:** Comprehend the concept, its causes, and its relevance in multiple regression analysis.
- **estimation amid Multicollinearity:** Understand the process and consequences of estimation in the presence of multicollinearity.
- **detection Techniques for Multicollinearity:** Learn how to identify multicollinearity using various methods like correlation matrices and variance inflation factors (VIF).
- **implications of Multicollinearity:** Analyze the potential effects and distortions that multicollinearity can introduce to statistical models and individual predictor variables.
- **applying Remedial Measures:** Learn to apply different strategies to mitigate multicollinearity, understanding the advantages and limitations of each approach.
- **critical Thinking and Problem-solving:** Develop skills to navigate challenges posed by multicollinearity in practical scenarios and to adapt strategies for effective management in personal statistical analyses.

5.3. Major Topics

- Nature of the Multicollinearity
- Estimation in the Presence of Multicollinearity
- Consequences of Multicollinearity
- Detection of Multicollinearity
- Remedial Measures

5.4. Summary of the Units

5.4.1. Nature of the Multicollinearity

The credit for the term "multicollinearity" goes to Ragnar Frisch. Originally, it referred to the presence of a "perfect" or exact linear relationship among one or more explanatory variables within a regression model. In the context of a k-variable regression involving explanatory variables X_1, X_2, ... , X_k (where $X_1 = 1$ for all observations to account for the intercept term), an exact linear relationship is considered to exist if the following condition is met:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0 \qquad (5.1)$$

Where $\lambda_1, \lambda_2, \ldots, \lambda_k$ are constants such that not all of them are zero simultaneously. This type of multicollinearity is called as perfect multicollinearity. There is a case where X variables are inter-correlated but not perfectly so,

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k + v_i = 0 \qquad (5.2)$$

Where v_i is a stochastic error term. Multicollinearity refers only to linear relationships among the X variable. Various factors can contribute to multicollinearity, as pointed out by Montgomery and Peck. These sources include:

- **Data collection method:** Multicollinearity may arise when data is collected over a limited range of values for the regressors in the population. For instance, if the sampling process is restricted to a specific segment of the variable's distribution.
- **Constraints in the model or population:** Physical constraints within the population being sampled can lead to multicollinearity. For example, in a regression of electricity consumption on income (X_2) and house size (X_3), families with higher incomes generally tend to have larger homes, introducing a correlation between these variables.
- **Model specification:** Multicollinearity can emerge due to the inclusion of polynomial terms in a regression model, particularly when the range of the X variable is narrow. This can lead to high correlations among the predictors.

- **Overdetermined model:** When a regression model has more explanatory variables than the number of observations available, it becomes overdetermined. In medical research, for instance, where data on many variables is collected from a small number of patients, multicollinearity can occur due to the scarcity of data points relative to the number of predictors.

By being aware of these potential sources of multicollinearity, researchers can better identify, and address issues related to collinearity in regression analysis.

5.4.2. Estimation in the Presence of Multicollinearity

In the case of perfect multicollinearity, the regression coefficients remain indeterminate, and their standard errors are infinite. In terms of the three-variable regression model, using the deviation form, where all the variables are expressed as deviations from their sample means, we can write the three-variable regression model as:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \quad (5.3)$$

Now from unit 3, we can write:

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \quad (5.4)$$

Assume that $X_{3i} = \lambda X_{2i}$, where λ is a nonzero constant. This implies that

$$x_{3i} = X_{3i} - \bar{X}_{3i} = \lambda X_{2i} - \lambda \bar{X}_{2i} = \lambda (X_{2i} - \bar{X}_{2i}) = \lambda x_{2i} \quad (5.5)$$

Substituting in the equation of $\hat{\beta}_2$, we will get:

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i} x_{2i})}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - (\lambda \sum x_{2i} x_{2i})^2} = \frac{0}{0} \quad (5.6)$$

Which is indeterminate expression. Similarly, it can be shown that $\hat{\beta}_3$ is also indeterminate. In essence, the situation implies that we cannot untangle the distinct impacts of X_2 and X_3 in the given sample. For all practical purposes, X_2 and X_3 become indistinguishable. This issue is particularly troublesome in applied econometrics, as the primary objective is precisely to isolate and assess the individual partial effects of each X on the dependent variable.

5.4.2.2 Estimation in the Presence of Imperfect Multicollinearity

Typically, there is no precise linear relationship among the X variables, particularly when dealing with economic time series data. Consequently, when examining the three-variable model in the deviation form presented in Eq. (5.3), instead of encountering exact multicollinearity, we may encounter another situation.

$$X_{3i} = \lambda X_{2i} + v_i \quad (5.7)$$

Here v_i is a stochastic variable with mean zero ($\bar{v}_i = 0$). This implies that:

$$x_{3i} = X_{3i} - \bar{X}_{3i} = (\lambda X_{2i} + v_i) - (\lambda \bar{X}_{2i})$$
$$= \lambda(X_{2i} - \bar{X}_{2i}) + v_i = \lambda x_{2i} + v_i \quad (5.8)$$

So the estimation of β_2 and β_3 is possible, with $\sum x_{2i} v_i = 0$, the expression for $\hat{\beta}_2$ can be written as:

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i} x_{2i})}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i} x_{2i})^2}$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i)(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2} \quad (5.9)$$

Similarly, we can obtain the estimated value of β_3.

5.4.3. Consequences of Multicollinearity

When near or high multicollinearity is present, several consequences are likely to arise:

5.4.3.1. Large Variances and Covariances of OLS Estimators

The Ordinary Least Squares (OLS) estimators, while Best Linear Unbiased Estimators (BLUE), exhibit large variances and covariances, making precise estimation challenging. Values of variances and covariances are given by:

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

$$cov(\hat{\beta}_3, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \quad (5.10)$$

Where r_{23} is the coefficient of correlation between X_2 and X_3. From Eq. (5.10), it is evident that as r_{23} approaches 1, the value of variances as well as the absolute value of the covariance between the two estimators also increases. The speed with which variances and covariances increase can be seen with the variance-inflating factor (VIF), which is defined as:

$$VIF = \frac{1}{(1 - r_{23}^2)} \quad (5.11)$$

VIF shows that how the variance of an estimator is inflated by the presence of multicollinearity. As r_{23}^2 approaches 1, the VIF approaches infinity. Using these equations for variances can be written as:

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2} VIF$$

$$var(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2} VIF \quad (5.12)$$

In case of k-variable model, variance of the kth coefficient can be expressed as:

$$var(\hat{\beta}_j) = \frac{\sigma^2}{\sum x_j^2\left(1-R_j^2\right)} = \frac{\sigma^2}{\sum x_j^2}\left(\frac{1}{1-R_j^2}\right) = \frac{\sigma^2}{\sum x_j^2} VIF_j \quad (5.13)$$

Where R_j^2 represents the R^2 in the regression of X_j on the remaining $(k-2)$ regressors and $\sum x_j^2 = \sum\left(X_j - \bar{X}_j\right)^2$. It should be noted that the inverse of the VIF is called tolerance (TOL). That is:

$$TOL_j = \frac{1}{VIF_j} = \left(1 - R_j^2\right) \quad (5.14)$$

When R_j^2 equals 1, indicating perfect collinearity, the tolerance factor (TOL_j) becomes 0. Conversely, when R_j^2 is 0, indicating no collinearity at all, TOL_j equals 1. Due to the close relationship between the VIF and the TOL, they can be used interchangeably in analyses.

5.4.3.2. Wider Confidence Interval

Due to the significant variance in the above consequence, the confidence intervals tend to be much wider, leading to a higher likelihood of accepting the "null hypothesis," where the true population coefficient is considered to be zero.

5.4.3.3. Insignificant "t" Ratios

The presence of significant variance can also result in the t ratio of one or more coefficients being statistically insignificant. This can be seen through the t-ratio [$t = \hat{\beta}/se(\hat{\beta})$]. Therefore, in such cases, one will increasingly accept the null hypothesis that the relevant true population value is zero.

5.4.3.4. A High R^2 but Few Significant t-Ratios

Despite the insignificance of the t ratio for one or more coefficients, the overall measure of goodness of fit, R^2, can be very high. Indeed, this is one of the signals of multicollinearity—insignificant t values but a high overall R^2 (and a significant F value)

5.4.3.5. Sensitivity of OLS Estimators and their Standard Errors to Small Change in Data

In case of imperfect multicollinearity, the OLS estimators and their standard errors become sensitive to small changes in the data, causing potential instability in the results.

5.4.4. Detection of Multicollinearity

As multicollinearity primarily stems from nonexperimental data prevalent in social sciences, it is a phenomenon inherent to the sample. Consequently, there is no

single definitive method to detect it or precisely measure its intensity. Instead, what we rely on are various rules of thumb, some of which are informal while others are more structured. In the following points, we will explore several of these rules to gain insights into identifying and dealing with multicollinearity in regression analysis.

5.4.4.1. A High R^2 but a Few Significant t-Ratios

If the R^2 value is high, perhaps over 0.8, it's likely that the F-test will usually dismiss the theory that the partial slope coefficients are all zero at the same time. However, the individual t-tests will reveal that none or only a small number of these partial slope coefficients are statistically distinct from zero. According to Kmenta (1986):

> *it is too strong in the sense that multicollinearity is considered as harmful only when all of the influences of the explanatory variables on Y cannot be disentangled.*

5.4.4.2. High Pair-Wise Correlations among Regressors

Another suggested rule of thumb is that if the pairwise or zero-order correlation coefficient between two regressors is high, say, more than 0.8, then multicollinearity is a serious problem. It should be noted that high zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even though the zero-order or simple correlations are comparatively low (say, less than 0.50).

5.4.4.3. Examination of Partial Correlations

Farrar and Glauber (1967) suggested using partial correlation coefficients rather than zero-order correlations. Thus, in the regression of Y on X_2, X_3, and X_4, a finding that $R^2_{1.234}$ is very high but $r^2_{12.3}$, $r^2_{13.24}$, and $r^2_{14.23}$ are comparatively low may suggest that the variables X_2, X_3, and X_4 are highly intercorrelated and that at least one of these variables is superfluous. On the other hand, Wichers (1975) has shown that the Farrar–Glauber partial correlation test is ineffective in that a given partial correlation may be compatible with different multicollinearity patterns.

5.4.4.4. Auxiliary Regressions

A method to identify which X variable correlates with other X variables involve running a regression of each X_i on the remaining X variables and calculating the associated R^2, referred to as R_i^2. Each of these regression analyses is known as an auxiliary regression, which is supplementary to the primary regression of Y on the X variables. In terms of F-test, it can be written as:

$$F_i = \frac{R^2_{X_i.X_2,X_3,\ldots,X_k}/(k-2)}{\left(1-R^2_{X_i.X_2,X_3,\ldots,X_k}\right)/(n-k+1)} \quad (5.15)$$

If the computed F exceeds the critical F_i at the chosen level of significance, it is taken to mean that the particular X_i is collinear with other X's; if it does not exceed the critical F_i, we say that it is not collinear with other X's, in which case we may retain that variable in the model.

Instead of formally testing all auxiliary R^2 values, one may adopt Klein' rule of thumb developed by Klein (1962), which suggests that multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2, that is, that obtained from the regression of Y on all the regressors.

5.4.4.5. Eigenvalues and Condition Index

From these eigenvalues, we can derive what is known as the condition number k defined as:

$$k = \frac{Maximum\ eigenvalue}{Minimum\ eigenvalue}$$

$$CI = \sqrt{\frac{Maximum\ eigenvalue}{Minimum\ eigenvalue}} = \sqrt{k} \quad (5.16)$$

Where presents the condition index. There's a general guideline to follow: if the value of k falls within the range of 100 to 1000, multicollinearity can be considered moderate to strong. If k surpasses 1000, the multicollinearity is deemed severe. In another perspective, if the Condition Index (CI, calculated as the square root of k) lies between 10 and 30, it indicates moderate to strong multicollinearity, whereas if it goes beyond 30, severe multicollinearity is inferred.

5.4.4.6. Tolerance and Variance Inflation Factor

Variance inflation factor (VIF) and tolerance (TOL) are presented in equations (5.11) and (5.14). As a rule of thumb, if the VIF of a variable exceeds 10, which will happen if R_j^2 exceeds 0.90, that variable is said be highly collinear (Kleinbaum, 1988).

5.4.4.7. Scatterplot

Observing scatterplot (as available in most of the software) is a good practice to see how the various variables in a regression model are related.

5.4.5. Remedial Measures

The "do nothing" school of thought is expressed by Blanchard (1967) as follows:

When students run their first ordinary least squares (OLS) regression, the first problem that they usually encounter is that of multicollinearity. Many of them conclude that there is something wrong with OLS; some resort to new and often creative techniques to get around the problem. But, we tell them, this is wrong. Multicollinearity is God's will, not a problem with OLS or statistical technique in general.

The following general guidelines can be attempted to resolve the issue of multicollinearity; however, their effectiveness is subject to the intensity of the collinearity issue.

5.4.5.1. A Priori Information

Consider the following model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (5.17)$$

Here, Y represents consumption, X_2 stands for income, and X_3 signifies wealth. As previously highlighted, income and wealth variables typically exhibit high collinearity. But let's assume we initially believe that β_3 equals $0.10\beta_2$; this means that the rate at which consumption changes in relation to wealth is one-tenth of the rate at which it changes in relation to income. Given this, we can conduct the subsequent regression:

$$Y_i = \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + u_i$$
$$Y_i = \beta_1 + \beta_2 (X_{2i} + 0.10 X_{3i}) + u_i$$
$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (5.18)$$

Once we obtain, $\hat{\beta}_2$, we can get $\hat{\beta}_3$ from the postulated relationship between β_2 and β_3. A priori information could come from previous empirical work in which the collinearity problem happens to be less serious or from the relevant theory underlying the field of study.

5.4.5.2. Combining Cross Sectional and Time Series Data

Pooling data is a mix of cross-sectional and time-series data to overcome the issue of multicollinearity. For instance, in studying the demand for automobiles in the US, using time series data might present a multicollinearity problem as price and income variables are usually highly collinear. An approach to tackle this issue, suggested by Tobin, involves using cross-sectional data, such as consumer panels or budget studies, to reliably estimate income elasticity, as prices do not vary significantly in these point-in-time data. The income-adjusted value of Y can then be used to estimate price elasticity.

However, this approach may create interpretational difficulties, as it presumes that the income elasticity estimated from cross-sectional data is the same as that derived

from time series analysis. Nevertheless, this technique can be quite valuable in scenarios where the cross-sectional estimates do not differ considerably from one cross section to another.

5.4.5.3. Dropping a Variable(s) and Specification Bias

Another simple method to handle serious multicollinearity is to remove one of the collinear variables. For example, in a model studying the relationship between consumption, income, and wealth, dropping the wealth variable could make the previously insignificant income variable become very significant.

However, removing a variable might introduce a specification error or bias, which occurs when the model used for analysis is incorrectly specified. For instance, if economic theory suggests that both income and wealth should be included in the consumption model, excluding wealth would introduce a specification bias. If we incorrectly exclude a variable from our model, our new estimate might either overestimate or underestimate the original variable's effect, leading to positive or negative bias, respectively.

Thus, dropping a variable from a model to resolve multicollinearity could introduce specification bias. In some cases, this "cure" might be worse than the problem, because while multicollinearity might hinder precise parameter estimation, omitting a variable might lead us astray regarding the true parameter values. It's important to remember that the estimators from Ordinary Least Squares (OLS) are best despite near collinearity.

5.4.5.4. Transformation of Variables

Transformation of variables can be useful to deal with multicollinearity, a common problem in time series data when variables, like income and wealth, tend to move in the same direction over time, resulting in a high correlation between them.

First Difference Form: The first method discussed is the 'first difference' approach. It involves running regression not on the original variables but on the differences between their successive values. This method often reduces multicollinearity as the levels of variables might be correlated, but their differences might not be. An added advantage of this approach is that it can make a non-stationary time series stationary, which is desirable because a stationary time series doesn't change its mean and variance systematically over time. However, this approach has its downsides: the new error term might not satisfy the assumption that disturbances are uncorrelated, and there is a loss of one observation due to the differencing process, reducing the degrees of freedom.

Ratio Transformation: The second method is the 'ratio transformation', which is particularly useful when variables grow over time and are likely to be correlated, as with GDP and population. To reduce collinearity, the model is expressed on a per capita basis, by dividing the entire equation by the population variable. However, the ratio model has its drawbacks too: if the original error term is homoscedastic (has constant variance), the new error term will be heteroscedastic (has variable variance).

In summary, while the first difference and ratio transformation methods can help alleviate multicollinearity, they come with their own set of potential issues. Therefore, care must be taken when deciding to use these transformations to resolve multicollinearity.

5.4.5.5. Additional or New data

Given that multicollinearity is a characteristic tied to a specific sample, it's plausible that a different sample, even if comprised of the same variables, might not face such severe collinearity as the first. At times, simply expanding the sample size (if feasible) could mitigate the issue of multicollinearity.

5.4.5.6. Reducing Collinearity in Polynomial Regressions

A special feature of polynomial regression models is that the explanatory variable(s) appear with various powers. Thus, the total cubic cost function involves the regression of total cost on output, $(\text{output})^2$, and $(\text{output})^3$. Although in practical scenarios it's often observed that expressing the explanatory variables as deviations from their mean value can significantly decrease multicollinearity, the issue might persist in some instances. In such cases, considering techniques like orthogonal polynomials may prove beneficial (Bradley & Srivastava, 1979).

5.4.5.7. Other Methods of Remedying Multicollinearity

Multivariate statistical techniques such as factor analysis and principal components or techniques such as ridge regression are often employed to “solve” the problem of multicollinearity.

5.5. Self-Assessment Questions

- Define multicollinearity and describe its causes in a multiple regression analysis.
- Why is multicollinearity a concern for regression models?
- What happens when we estimate coefficients in the presence of multicollinearity?
- Can a model with multicollinearity provide reliable estimates? Why or why not?
- Describe two methods that can be used to detect multicollinearity.
- Explain how variance inflation factors (VIF) can be used to identify multicollinearity.
- How does multicollinearity affect the interpretability and reliability of regression coefficients?
- Can multicollinearity lead to overfitting in a model? Explain.
- Describe two remedial measures that can be used to mitigate multicollinearity.
- What are the potential downsides of these remedial measures?
- Imagine you are working with a dataset that has severe multicollinearity. How would you approach the problem?
- Can multicollinearity always be avoided or resolved? Justify your answer.

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Blanchard, O. J. (1967) Comment, Journal of Business and Economic Statistics, 5, 449–451.
- Bradley, R. A., & Srivastava, S. S. (1979). Correlation and Polynomial Regression. American Statistician, 33, pp. 11–14.
- Farrar, D. E. & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited, Review of Economics and Statistics, 49, 92–107.
- Griffiths, W. E., Hill, R. C., & Judge, G. G. *Learning and Practicing Econometrics*, Wiley.
- Klein, L. R. An Introduction to Econometrics, Prentice-Hall, Englewood Cliffs, NJ, p. 101.
- Kleinbaum, D. G., Kupper, L. L., & and Muller, K. E. Applied Regression Analysis and Other Multivariate Methods, Latest edition., PWS-Kent, Boston, Mass., p. 210.
- Kmenta, J. *Elements of Econometrics*, Latest edition, Macmillan, New York, p. 431.
- Wichers, C. R. (1975). The Detection of Multicollinearity: A Comment, Review of Economics and Statistics, 57, 365–366.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 06

HETEROSCEDASTICITY

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

6.1. INTRODUCTION

In this comprehensive Unit, we delve into the multifaceted concept of heteroscedasticity, a phenomenon that plays a crucial role in regression analysis and econometric modeling. The chapter is meticulously structured to provide a holistic understanding of the subject, beginning with an exploration of the very nature of heteroscedasticity. This foundational section sets the stage for a detailed examination of the various methods to detect heteroscedasticity, both informal and formal.

The detection section is further subdivided to cover a range of formal tests, including the Park Test, Glejser Test, Spearman's Rank Correlation Test, Goldfeld-Quandt Test, Breusch-Pagan-Godfrey Test, and White's General Heteroscedasticity Test. Each of these tests is elucidated with precision, offering insights into their applications, advantages, and limitations. Following the detection, the Unit transitions into an analysis of the consequences of heteroscedasticity, particularly when it is unaddressed in regression models. This leads to the final substantive section, which presents various solutions to heteroscedasticity problems, offering both theoretical frameworks and practical tools to address this complex issue.

6.2. OBJECTIVES

The possible objectives for students studying this material could include:

- **understanding the Nature of Heteroscedasticity:** To comprehend the underlying concept and characteristics of heteroscedasticity in regression models.
- **learning Detection Techniques:** To acquire skills in detecting heteroscedasticity using both informal and formal methods, including understanding the underlying principles and applications of various tests.
- **analyzing Specific Formal Tests:** To delve into the specifics of formal tests such as the Park Test, Glejser Test, Spearman's Rank Correlation Test, Goldfeld-Quandt Test, Breusch-Pagan-Godfrey Test, and White's General Heteroscedasticity Test, understanding their equations, advantages, and disadvantages.
- **evaluating the Consequences of Heteroscedasticity:** To analyze the impact of heteroscedasticity on regression analysis, particularly when using Ordinary Least Squares (OLS), and to understand the implications for statistical inference.

- **implementing Solutions to Heteroscedasticity Problems:** To learn and apply various remedial measures for heteroscedasticity, including both theoretical frameworks and practical tools.
- **critical Thinking and Application:** To apply the concepts learned to real-world data and scenarios, recognizing the presence of heteroscedasticity, and choosing appropriate detection methods and remedies.
- **self-Assessment and Reflection:** To engage in self-assessment through provided questions, reflecting on understanding and identifying areas for further exploration.
- **exploring Additional Resources:** To utilize the provided textbooks, supplies, and additional readings for deeper exploration and mastery of the subject.
- **ethical Consideration and Best Practices:** To understand the ethical considerations and best practices in applying these methods, recognizing the potential limitations and biases.
- **interdisciplinary Integration:** To recognize the interdisciplinary nature of heteroscedasticity, understanding its relevance and application in economics, finance, social sciences, and other related fields.

These objectives align with the Unit's comprehensive approach to heteroscedasticity, providing students with a well-rounded understanding of the subject, from foundational concepts to advanced applications.

6.3. Major Topics

- Nature of the Heteroscedasticity
- Detection of Heteroscedasticity
- Consequences of Heteroscedasticity
- Solutions to Heteroscedasticity Problems

6.4. Summary of the Units

6.4.1. Nature of the Heteroscedasticity

A key presumption of the classical linear regression model is that given any selected values of the explanatory variables, the variance of each error term u_i is a fixed quantity, denoted by σ^2. This principle is known as homoscedasticity, which essentially means equal (homo) dispersion (scedasticity), or in other words, equal variance. This can be represented symbolically as:

$$E(u_i)^2 = \sigma^2 \qquad i = 1,2,3, \dots, n \tag{6.1}$$

In contrast, if the conditional variance of Y_i increases as X are not the same. Hence, there is heteroscedasticity. Symbolically,

$$E(u_i)^2 = \sigma_i^2 \tag{6.2}$$

Notice the subscript of σ^2, which reminds us that the conditional variances of u_i (= conditional variances of Y_i) are no longer constant. There are several reasons why the variances of u_i may be variable, some of which are as follows:

- **Measurement Errors:** The variance of the error term in a regression model may increase as a result of measurement errors in the independent variables. This is particularly likely if the measurement errors increase with the size of the variable being measured.
- **Structural Changes in the Economy:** Changes in the structure of the economy over time can cause the variance of the error term to change. For example, if the economy is growing, the variance of the error term may increase over time.
- **Changes in Technology:** Technological changes can cause the variance of the error term to change. For example, if a new technology is introduced, it may cause the variance of the error term to increase.
- **Changes in Policy:** Changes in government policy can cause the variance of the error term to change. For example, if the government introduces a new policy, it may cause the variance of the error term to increase.

- **Data Collection Errors:** Errors in data collection can contribute to variable variances in the error term. Erroneous data, whether due to measurement inaccuracies or other inconsistencies, can lead to a distortion in the error term's variance. These errors may be random or systematic, and their impact on the error term's variance can be multifaceted and profound.
- **Changes in the Distribution of Income:** Changes in the distribution of income can cause the variance of the error term to change. For example, if income inequality increases, it may cause the variance of the error term to increase.
- **Changes in the Variability of Explanatory Variables:** Changes in the variability of the explanatory variables can cause the variance of the error term to change. For example, if the variability of the explanatory variables increases, it may cause the variance of the error term to increase.
- **Specification Errors:** If the functional form of the regression model is incorrectly specified, it can cause the variance of the error term to change. For example, if a linear model is used when the true relationship is nonlinear, it can cause the variance of the error term to increase.

In summation, the variability in the variances of the error term is a multifaceted issue, stemming from heteroscedasticity, model misspecification, data collection errors, and specific economic factors. The interplay of these elements creates a complex landscape that requires careful consideration and analysis.

6.4.2. Detection of Heteroscedasticity

There can be informal and formal methods to detect the possible presence of heteroscedasticity.

6.4.2.1. Informal Methods

Graphical methods are a popular and intuitive way to detect heteroscedasticity in a dataset. They allow for a visual inspection of the data, which can often reveal patterns or inconsistencies that might not be apparent through statistical tests alone.

Graphical methods serve as an indispensable tool in the detection of heteroscedasticity, a phenomenon that refers to the unequal scatter of residuals across the range of fitted values in a regression model. These methods primarily revolve around visual inspection, providing an intuitive and accessible approach to understanding the underlying patterns in the data.

Firstly, scatter plots are employed to visually inspect the relationship between residuals and predicted values or independent variables. By plotting these variables against each other, one can discern whether the variance of the residuals remains constant. A funnel-shaped pattern or any discernible non-random pattern may be indicative of heteroscedasticity. This method, although simple and intuitive, is inherently subjective and may require corroboration through more definitive tests.

A seminal study by Cook and Weisberg titled "Diagnostics for heteroscedasticity in regression" (1983) provides a diagnostic test for heteroscedasticity based on the score statistic and presents a graphical procedure to complement the score test. They emphasize the importance of both graphical and non-graphical procedures, with the usual graphical procedure consisting of plotting the ordinary least squares residuals against fitted values or an explanatory variable.

Secondly, the Residual vs. Fitted Values Plot, a specific type of scatter plot, is utilized to detect non-constant variance in the residuals. This plot juxtaposes the residuals from the regression model against the fitted values, allowing for a direct visual check for one of the key assumptions of linear regression. A random scatter around zero signifies homoscedasticity, while any discernible pattern may signal heteroscedasticity.

In conclusion, graphical methods, though exploratory in nature, offer a valuable starting point in the analysis of heteroscedasticity. They bridge the gap between complex statistical concepts and intuitive visual understanding, providing a multifaceted perspective on the data. However, their subjective nature necessitates the use of complementary statistical tests to arrive at a definitive conclusion regarding the presence of heteroscedasticity.

6.4.2.2. Formal Methods

6.4.2.2.1. Park Test

The Park test, named after R. E. Park, is a formal method used to detect heteroscedasticity in regression models. The Park test formalizes the graphical method by suggesting that the variance of the error term σ_i^2 is a function of the explanatory variable X_i. The functional form suggested $\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$, where v_i is the stochastic disturbance term. Following are the steps:

- **Step 1 - Estimate the Original Model:** Run the OLS regression disregarding the heteroscedasticity question and obtain residuals $\hat{u}_i$.
- **Step 2 – Assume Error Variance and Run the Regression:** Assume that the error variance is related to the explanatory variable.

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i \quad (6.3)$$

Further, run the regression using the above equation.

- **Step 3 – Interpret the Results:** If β turns out to be statistically significant, it would suggest that heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity.

Goldfeld and Quandt (1972) have posited that the error term within the test's equation may itself be heteroscedastic, undermining the OLS assumptions. Additionally, the test's reliance on a specific functional form for the relationship between error variance and the explanatory variable may not always hold true, adding a layer of complexity and potential limitation to its application. Thus, while the Park test provides a valuable tool in certain scenarios, its nuanced intricacies necessitate careful consideration in empirical research.

The Park test, as described in Park (1966), provides a formalized approach to detecting heteroscedasticity. While it offers an empirically appealing and exploratory method, it also comes with some recondite challenges, such as potential heteroscedasticity in the error term and assumptions about the functional form. These complexities necessitate careful consideration and application of the test in empirical research.

6.4.2.2.2. Glejser Test

The Glejser test is another method used to detect heteroscedasticity in regression models. Here's a detailed description of the test, including the steps, equations, and a short paragraph on the advantages and disadvantages, as described in the provided document:

- **Step 1 – Estimate the Original Model:** Run the OLS regression disregarding the heteroscedasticity question $Y_i = \beta_1 + \beta_2 X_i + u_i$ and obtain residuals $\hat{u}_i$.
- **Step 2 – Regress Absolute Values of Residuals on Explanatory Variable:** Glejser suggests regressing the absolute values of *X* variable that is suspected to be the cause of heteroscedasticity.

  $$|\hat{u}_i| = \alpha_1 + \alpha_2 X_i + v_i \quad (6.4)$$

  Where v_i is the error term of this regression. Glejser also proposed following functional forms, out of which anyone can be used:

  $$|\hat{u}_i| = \alpha_1 + \alpha_2 \sqrt{X_i} + v_i \quad (6.5)$$

  $$|\hat{u}_i| = \alpha_1 + \alpha_2 \frac{1}{X_i} + v_i \quad (6.6)$$

$$|\hat{u}_i| = \alpha_1 + \alpha_2 \frac{1}{\sqrt{X_i}} + v_i \quad (6.7)$$

$$|\hat{u}_i| = \sqrt{\alpha_1 + \alpha_2 X_i} + v_i \quad (6.8)$$

$$|\hat{u}_i| = \sqrt{\alpha_1 + \alpha_2 X_i^2} + v_i \quad (6.9)$$

- **Step 3 – Interpret the Results:** If the coefficient of X_i is statistically significant, it suggests the presence of heteroscedasticity.

The Glejser test, akin in spirit to the Park test, offers a method to detect heteroscedasticity by focusing on the relationship between the absolute values of residuals and the suspected explanatory variable. One of its advantages is its applicability to large samples, providing generally satisfactory results. It may also be used in small samples as a qualitative device to learn about heteroscedasticity. However, the test's simplicity may also be seen as a limitation, as it does not provide a comprehensive understanding of the underlying structure of heteroscedasticity. The test's reliance on the specific form of the relationship between the residuals and the explanatory variable may also pose challenges in some applications.

The Glejser test provides a straightforward and practical approach to detecting heteroscedasticity, especially in large samples. While it offers an accessible method, its recondite limitations and assumptions necessitate careful consideration in empirical research. The test's development and application are well-documented in studies by Glejser (1969).

6.4.2.2.3. Spearman's Rank Correlation Test

The Spearman's Rank Correlation Test is a statistical method used to detect heteroscedasticity in regression models.

- **Step 1 – Estimate the Original Model:** Run the OLS regression disregarding the heteroscedasticity question $Y_i = \beta_1 + \beta_2 X_i + u_i$ and obtain residuals $\hat{u}_i$.
- **Step 2 – Rank the Absolute Values of Residuals:** Ignoring the sign of $\hat{u}_i$, take their absolute value $|\hat{u}_i|$, rank both $|\hat{u}_i|$ and X_i (or $\hat{Y}_i$) according to an ascending or descending order. By using these ranks compute Spearman's Rank Correlation Coefficient.
- **Step 3 – Test the Significance:** Assuming that the population rank correlation coefficient ρ_s is zero and $n > 8$, the significance of the sample r_S can be tested by the t test as follows:

$$t = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \quad (6.10)$$

With df $n-2$.

- **Step 4 – Interpret the Results:** If the computed t value exceeds the critical t value, heteroscedasticity may be accepted; otherwise, it may be rejected.

The Spearman's Rank Correlation Test offers a non-parametric approach to detecting heteroscedasticity, making it applicable without assuming a specific distribution for the data. It provides a simple and intuitive method that can be particularly useful in exploratory data analysis. However, the test's simplicity may also be seen as a limitation, as it does not provide a comprehensive understanding of the underlying structure of heteroscedasticity. Moreover, the test's reliance on ranking may pose challenges in some applications, especially when dealing with tied ranks or small sample sizes.

The Spearman's Rank Correlation Test, as a non-parametric method, provides a straightforward approach to detecting heteroscedasticity. While it offers an accessible method, its recondite limitations and assumptions necessitate careful consideration in empirical research. The test's development and application are well-documented in statistical literature and can be a valuable tool in econometric analysis.

6.4.2.2.4. Goldfeld-Quandt Test

The Goldfeld-Quandt Test is a widely used method to detect heteroscedasticity in regression models.

- **Step 1 – Divide the Data:** Order the data based on the explanatory variable suspected to be related to the heteroscedastic variance. Omit a certain number of central observations (c) and divide the remaining data ($n-c$) into two groups each having $(n-c)/2$ observations.

- **Step 2 – Estimate two Separate Regressions and Compute RSS:** Estimate two separate regressions for the two groups.

  $$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \quad (6.11)$$

  Compute the residual sum of squares (RSS) for each group. RSS_1 representing the RSS from the regression corresponding to the smaller X_i values and RSS_2 that from the larger X_i values.

- **Step 3 – Compute the Test Statistic:** The test statistic is the ratio of the two RSS values.

$$F = \lambda = \frac{RSS_2/\mathrm{df}}{RSS_1/\mathrm{df}} \quad (6.12)$$

These RSS each have $[(n-c)/2]-k$ degree of freedom, where k is the number of parameters to be estimated, including the intercept.

- **Step 4 – Interpret the Results:** If the computed F value exceeds the critical F value at the chosen level of significance, heteroscedasticity may be accepted; otherwise, it may be rejected.

The Goldfeld-Quandt Test offers a simple and intuitive method to detect heteroscedasticity by focusing on the relationship between the variance of the error term and one of the explanatory variables. One of its advantages is its applicability to various types of regression models and its ability to sharpen the difference between small and large variance groups. However, the test's effectiveness depends on how the central observations are omitted, and choosing the wrong number can diminish the power of the test. Moreover, the test requires reordering the observations, which may pose challenges in some applications.

The Goldfeld-Quandt Test, developed by Stephen Goldfeld and Richard Quandt, provides a practical approach to detecting heteroscedasticity. While it offers an accessible method, its recondite limitations and assumptions necessitate careful consideration in empirical research. The test's development and application are well-documented in econometric literature, including Monte Carlo experiments done by Goldfeld and Quandt, and it remains a valuable tool in econometric analysis.

6.4.2.2.5. Breusch-Pagan-Godfrey Test

The Breusch-Pagan-Godfrey test is a combination of methods introduced by Trevor Breusch, Adrian Pagan, and L. Godfrey. It's designed to detect heteroscedasticity, specifically whether the variances of the error terms in a regression model are constant or not.

The test procedure is explained in the following steps:

- **Step 1 - Estimate the Original Model:** Fit the original regression model and obtain the residuals $\hat{u}_i$.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i \quad (6.13)$$

- **Step 2 - Calculate the Squared Residuals:** Square the residuals to obtain a measure of the variance.

$$\tilde{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n} \quad (6.14)$$

This is maximum likelihood estimator (ML) of σ^2.

- **Step 3 – Calculate P_i:** Construct the variable P_i defined as:
$$P_i = \frac{\hat{u}_i^2}{\tilde{\sigma}^2} \quad (6.15)$$

- **Step 4 – Regress P_i on Z's**: Regress the constructed P_i from previous step on Z's.
$$P_i = \alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \cdots + \alpha_m Z_{mi} + v_i \quad (6.16)$$
Where v_i is the residual term of this regression.

- **Step 5 – Calculate Test Statistic:** The test statistic is computed using the ESS (explained sum of squares) from the regression.
$$\Theta = \frac{1}{2}(ESS) \quad (6.17)$$
Assuming u_i are normally distributed, Θ follows the chi-square distribution with $(m - 1)$ degree of freedom.

- **Step 6 – Compare with Chi-Squared Distribution:** Compare the test statistic with the critical value from the chi-squared distribution to decide. Therefore, if in an application the computed Θ $(= \chi^2)$ exceeds the critical χ^2 value at the chosen level of significance, one can reject the hypothesis of homoscedasticity; otherwise, one does not reject it.

The Breusch-Pagan-Godfrey test, a prominent tool for detecting heteroscedasticity, offers several advantages. It avoids some of the limitations found in other tests, such as the Goldfeld-Quandt test, by providing a more flexible approach that does not rely solely on identifying the correct variable with which to order the observations. This flexibility extends to its applicability across various types of regression models, not just linear ones, making it a versatile tool. Moreover, the BPG test is often more sensitive to heteroscedasticity, enhancing its power in identifying non-constant variance in error terms. However, this test is not without its disadvantages. It is sensitive to the assumption of normality, which can affect its power and validity if the error terms are not normally distributed. The complexity in choosing the test statistic might make it more challenging to understand and implement. Furthermore, as a large-sample test, its application in small samples may not be strictly justified, and other tests may be statistically more powerful in both large and small samples. These nuances reflect the intricate balance of strengths and potential challenges associated with the Breusch-Pagan-Godfrey test in the field of econometrics.

The Breusch-Pagan-Godfrey test offers a nuanced approach to detecting heteroscedasticity, balancing sensitivity, and flexibility. However, its sensitivity to normality and complexity in computation may pose challenges. The test's

development and application are well-documented in studies by Breusch and Pagan (1979) and Godfrey (1978).

6.4.2.2.6. White's General Heteroscedasticity Test

White's General Heteroscedasticity Test is a well-known method used to detect heteroscedasticity in regression models. White's General Heteroscedasticity Test, developed by White (1980). The White test can be a test of (pure) heteroscedasticity or specification error or both. Here's a detailed description including the steps are as follow:

- **Step 1 - Estimate the Original Model:** Fit the original regression model and obtain the residuals $\hat{u}_i$.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (6.18)$$

- **Step 2 – Regress Squared Residuals on Explanatory Variables:** White suggests regressing the squared residuals on the original explanatory variables, their squared terms, and cross-product terms.

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i \quad (6.19)$$

  Obtain the R^2 from this auxiliary regression.

  Step 3 – Compute Test Statistic: The test statistic is R^2 $\left(nR^2 \sim \chi^2_{\text{df}}\right)$, where n is the sample size and R^2 is the coefficient of determination from previous step.

- **Step 4 – Interpret the Results:** If the computed value exceeds the critical chi-square value at the chosen level of significance, heteroscedasticity may be accepted; otherwise, it may be rejected. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$.

If a model has several regressors, then introducing all the regressors, their squared terms, and their cross products can quickly consume degrees of freedom. In this case, the test can be modified to conserve degrees of freedom and use only explanatory variables and their squared terms in the auxiliary regression.

White's General Heteroscedasticity Test offers a robust method to detect heteroscedasticity without relying on the normality assumption, making it versatile and easy to implement. One of its advantages is its ability to test for both heteroscedasticity and specification error. However, the test's effectiveness may be limited by its consumption of degrees of freedom, especially when introducing all the regressors, their squared terms, and cross products. Additionally, the test may

have low power against alternatives and is of little help in identifying the factors or variables that cause heteroscedasticity.

6.4.3. Consequences of Heteroscedasticity

The consequences of using Ordinary Least Squares (OLS) in the presence of heteroscedasticity are multifaceted and can have significant implications for statistical inference. Here's a detailed exploration of these consequences:

- **Inefficiency of OLS Estimators:** In the presence of heteroscedasticity, OLS estimators remain linear and unbiased but lose their efficiency (minimum variance property). This means that there may be other estimators that provide a more accurate estimate of the population parameters.
- **Overestimation of Standard Errors:** The usual OLS standard errors are either too large (for the intercept) or generally too small (for the slope coefficient) in relation to those obtained by OLS allowing for heteroscedasticity. This inconsistency overestimates the true standard error obtained by the Generalized Least Squares (GLS) procedure.
- **Invalidity of Conventional t and F Tests:** If heteroscedasticity is present, the conventional t and F tests become invalid. This is because the variance formula used in these tests does not take into account the non-constant variance of the error term, leading to inaccurate results.
- **Difficulty in Applying GLS:** Although GLS is superior in the presence of heteroscedasticity, it is not always easy to apply in practice. Unless heteroscedasticity is very severe, one may not abandon OLS in favor of GLS or Weighted Least Squares (WLS).
- **Potential Misinterpretation of Coefficients:** Confidence intervals based on OLS that do not account for heteroscedasticity will be unnecessarily larger. As a result, what appears to be a statistically insignificant coefficient may, in fact, be significant if the correct confidence intervals were established based on the GLS procedure.
- **Challenges in Correction:** Even if heteroscedasticity is suspected and detected, it is not easy to correct the problem. If the sample is large, one can obtain White's heteroscedasticity-corrected standard errors of OLS estimators, but otherwise, it may require educated guesses of the likely pattern of heteroscedasticity.

The presence of heteroscedasticity in a regression model poses a serious problem, particularly when using OLS. The recondite nature of these consequences

necessitates a careful approach to model specification and testing. In short, if we persist in using the usual testing procedures despite heteroscedasticity, whatever conclusions we draw or inferences we make may be very misleading. The message is clear: In the presence of heteroscedasticity, it is advisable to use methods like GLS that explicitly account for the non-constant variance of the error term. However, the practical challenges in applying these methods and the nuanced intricacies of the consequences mean that empirical researchers must exercise caution and consider the specific context of their analysis.

6.4.4. Solutions to Heteroscedasticity Problems

The remedial measures for heteroscedasticity are essential to ensure the efficiency and validity of the estimators in regression analysis. Here's a detailed exploration of these measures:

6.4.4.1. When σ_i^2 is Known

If σ_i^2 is known then the most significant method of correcting heteroscedasticity is by means of Weighted Least Squares (WLS), for the estimators thus obtained are BLUE. This method involves transforming the original data by dividing each observation by the known standard deviation of the error term. This creates a new model where the error term has a constant variance.

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{u_i}{\sigma_i} \qquad (6.20)$$

6.4.4.1. When σ_i^2 is Not Known

6.4.4.1.1. White's Heteroscedasticity-Corrected Standard Errors

In large samples, one can obtain White's heteroscedasticity-corrected standard errors of OLS estimators and conduct statistical inference based on these standard errors. Incidentally, White's heteroscedasticity corrected standard errors are also known as robust standard errors.

6.4.4.1.2. Plausible Assumptions about Heteroscedasticity Pattern

Based on OLS residuals, one can make educated guesses of the likely pattern of heteroscedasticity and transform the original data in such a way that in the transformed data there is no heteroscedasticity. Several plausible assumptions about the heteroscedasticity are as follow:

- **Assumption 1:** The error variance is proportional to X_i^2.

$$E(U_i^2) = \sigma^2 X_i^2 \qquad (6.21)$$

one may transform the original model as follows.

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 \frac{X_i}{X_i} + \frac{u_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + v_i \quad (6.22)$$

- **Assumption 2:** The error variance is proportional to X_i.

$$E(U_i^2) = \sigma^2 X_i \quad (6.23)$$

then the original model can be transformed as follows:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2 \frac{X_i}{\sqrt{X_{i_i}}} + \frac{u_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + v_i \quad (6.24)$$

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

$$E(U_i^2) = \sigma^2 [E(Y_i)]^2 \quad (6.25)$$

Therefore, if we transform the original equation as follows:

$$\frac{Y_i}{E(Y_i)} = \frac{\beta_1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)} = \beta_1 \frac{1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + v_i \quad (6.26)$$

- **Assumption 4:** A log transformation such as

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad (6.27)$$

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 scales in which the variables are measured, thereby reducing a tenfold difference between two values to a twofold difference.

The remedial measures for heteroscedasticity are multifaceted and require careful consideration of the specific context of the analysis. While methods like Weighted Least Squares and White's heteroscedasticity-corrected standard errors offer robust solutions, the recondite nature of these measures and the potential for overreaction necessitate a nuanced approach. The choice of remedy should be guided by the severity of the heteroscedasticity, the known or unknown nature of the error variance, and the overall fit and validity of the model.

These insights align with the broader econometric literature, emphasizing the importance of understanding the underlying structure of the data and the specific assumptions of the chosen model. The remedial measures for heteroscedasticity are not one-size-fits-all solutions but rather tools that can be adapted and applied as needed to ensure the integrity and interpretability of regression analysis.

6.5. Self-Assessment Questions

- Define heteroscedasticity and explain how it differs from homoscedasticity.
- What are the common causes of heteroscedasticity in regression models?
- Describe the difference between informal and formal methods of detecting heteroscedasticity.
- How does the Spearman's Rank Correlation Test work, and in what situations is it most applicable?
- Compare and contrast the Glejser Test and the Goldfeld-Quandt Test. What are the key advantages and disadvantages of each?
- Explain the underlying principles of White's General Heteroscedasticity Test. How does it differ from the Breusch-Pagan-Godfrey Test?
- What are the main consequences of using OLS in the presence of heteroscedasticity? Provide examples.
- How does heteroscedasticity affect the efficiency and validity of OLS estimators?
- Describe the method of Weighted Least Squares (WLS) and explain how it can be used to correct heteroscedasticity?
- What is White's heteroscedasticity-corrected standard errors, and when are they most appropriately used?
- Imagine you are working with a dataset that exhibits heteroscedasticity. Outline the steps you would take to detect and correct this issue.
- Critically evaluate the statement: "Heteroscedasticity has never been a reason to throw out an otherwise good model." Provide evidence from the chapter to support your answer.
- Select one of the additional readings provided in the chapter and summarize its main findings or arguments related to heteroscedasticity.
- How would you apply the concepts learned in this unit to a real-world scenario in your field of interest?

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Breusch, T., & Pagan, A. (1979). A simple test for Heteroscedasticity and random coefficient variation, *Econometrica*, 47, 1287-1294.
- Cook, R. D., & Weisberg, S. (1983). *Diagnostics for Heteroscedasticity in Regression*, Technical Report No. 405, Department of Applied Statistics, University of Minnesota.
- Glejser, H. (1969). A new test for Heteroskedasticity. *Journal of the American Statistical Association*, 64(325), pp. 316-323.
- Godfrey, L. (1978). Testing for Multiplicative Heteroscedasticity, *Journal of Econometrics*, 8, 227-236.
- Goldfeld, S. M., & Quandt, R. E. *Nonlinear Methods in Econometrics*, North Holland Publishing Company, Amsterdam.
- Griffiths, W. E., Hill, R. C., & Judge, G. G. *Learning and Practicing Econometrics*, Wiley.
- Kmenta, J. *Elements of Econometrics*, Latest edition, Macmillan, New York, p. 431.
- Park, R. E. (1966). Estimation with Heteroscedastic Error Terms. *Econometrica*, 34(4), p. 888.
- White, H. (1980). A heteroscedasticity consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, 48, 817-818.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 07

AUTOCORRELATION

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

7.1. INTRODUCTION

Autocorrelation is a prevalent phenomenon in time-series data and is the central focus of this Unit. The exploration begins by elucidating the nature of autocorrelation, defining what it means, and how it manifests within various data sets. Understanding this foundational concept sets the stage for a deeper examination of its consequences, revealing how autocorrelation can impact statistical inference and the interpretation of models. The unit then delves into various methods for detecting autocorrelation, providing both visual and statistical tools. The graphical method offers an intuitive way to visualize autocorrelation, while formal tests such as the Runs Test, Durbin-Watson d test, and Breusch–Godfrey test provide rigorous statistical means to detect it.

Recognizing that detection is only part of the solution, the unit also covers remedial measures for autocorrelation. This section provides strategies for addressing the issue, both when the autocorrelation parameter (ρ) is known and when it is unknown. Various techniques are explored, including the First Difference Method, methods based on Durbin-Watson d statistics, estimating ρ from the residuals, and iterative methods of estimating ρ.

Overall, this unit offers a comprehensive examination of autocorrelation, from its fundamental nature to advanced detection and remediation techniques. It serves as a vital resource for researchers, analysts, and students seeking to understand and address autocorrelation in their data analysis endeavors. Whether you are new to the subject or looking to deepen your understanding, this unit provides the insights and tools needed to navigate the complex landscape of autocorrelation.

7.2. OBJECTIVES

The objectives for students studying the unit on autocorrelation are designed to provide a comprehensive understanding of this complex statistical concept. In the end of the unit, students should be able to:

- **understand the Nature of Autocorrelation:** Define autocorrelation and explain how it manifests in time-series data. Recognize the underlying patterns that may lead to autocorrelation.
- **identify the Consequences of Autocorrelation:** Analyze how autocorrelation can affect statistical inference and model interpretation. Understand the potential biases and inefficiencies it may introduce.
- **detect Autocorrelation through Various Methods:** Utilize both graphical

and formal statistical tests to detect autocorrelation. Apply methods such as the Runs Test, Durbin-Watson d test, and Breusch–Godfrey test.

- **implement Remedial Measures:** Develop strategies to address autocorrelation, considering both known and unknown autocorrelation parameters (ρ). Explore techniques like the First Difference Method, Durbin-Watson d statistics, and iterative methods.
- **apply Practical Solutions:** Translate theoretical understanding into practical application. Demonstrate the ability to diagnose and remedy autocorrelation in real-world data sets.
- **evaluate Different Approaches:** Compare and contrast various detection and remediation techniques, recognizing the advantages and disadvantages of each.
- **cultivate Critical Thinking:** Encourage critical evaluation of autocorrelation, fostering the ability to question assumptions, interpret results, and make informed decisions in data analysis.
- **integrate Knowledge with Technology:** Leverage modern computing tools to efficiently implement detection and remediation methods, enhancing practical skills in data analysis.
- **foster Lifelong Learning:** Instill a curiosity and willingness to continue exploring autocorrelation and related statistical concepts, recognizing the evolving nature of the field.

These objectives align with the unit's comprehensive examination of autocorrelation, aiming to equip students with the knowledge, skills, and critical thinking necessary to effectively understand and address this complex statistical phenomenon.

7.3. Major Topics

- Nature of The Autocorrelation
- Consequences of Autocorrelation
- Methods of detection of Autocorrelation
- Remedial Measures

7.4. Summary of the Units

7.4.1. Nature of the Autocorrelation

Autocorrelation, also known as serial correlation, refers to the correlation between members of a series of observations ordered in time (as in time series data) or space (as in cross-sectional data). In the context of regression, the classical linear regression model assumes that autocorrelation does not exist in the disturbances u_i. Symbolically, the absence of autocorrelation is represented in the classical linear regression model's assumptions:

$$cov(u_i, u_j | X_i, X_j) = E(u_i u_j) = 0 \qquad i \neq j \qquad (7.1)$$

However, if there is such a dependence, we have autocorrelation. Symbolically,

$$E(u_i u_j) \neq 0 \qquad i \neq j \qquad (7.2)$$

In this situation, the disruption caused by a strike this quarter may very well affect output next quarter, or the increases in the consumption expenditure of one family may very well prompt another family to increase its consumption expenditure.

There are several reasons of serial correlation, some of which are as follows:

- **Inertia:** Inertia refers to the sluggishness or cycles observed in most economic time series such as GNP, price indexes, production, employment, and unemployment. These cycles can create patterns, leading to autocorrelation.
- **Specification Bias: Excluded Variables Case or Incorrect Functional Form:** Sometimes, patterns in residuals are observed because the model is mis-specified. This could be due to the exclusion of some essential variables or incorrect functional form. A simple test of this would be to run different models and see whether autocorrelation disappears when the correct model is run.
- **Cobweb Phenomenon:** The Cobweb Phenomenon is a common occurrence in the supply of many agricultural commodities. It describes a situation where the supply reacts to price changes with a delay, typically one time

period. This lag happens because supply decisions, such as planting crops, take time to implement due to the gestation period. For example, farmers' decisions at the beginning of the year for planting crops are influenced by the prices from the previous year. This relationship can be represented by the equation:

$$Supply_t = \beta_1 + \beta_2 P_{t-1} + u_t \quad (7.3)$$

If the price at the end of period t, denoted as P_t, turns out to be lower than the price in the previous period P_{t-1}, farmers may decide to produce less in the next period $t + 1$. This leads to a situation where disturbances u_t are not random. If farmers overproduce in year t, they are likely to reduce their production in $t + 1$, and this pattern continues, creating a cobweb-like pattern.

In essence, the Cobweb Phenomenon illustrates a cyclical pattern in agricultural supply, driven by delayed reactions to price changes. This pattern can lead to non-random disturbances in supply, reflecting the strategic adjustments made by farmers in response to price fluctuations.

- **Lags of Variables:** In time series regression, lags can be a significant source of autocorrelation, particularly when modeling relationships like consumption expenditure on income. It's common to find that current consumption expenditure depends not only on current income but also on the consumption expenditure of the previous period. This relationship can be expressed as:

$$Consumption_t = \beta_1 + \beta_2 income + \beta_3 consumption_{t-1} + u_t \quad (7.4)$$

This type of regression is referred to as autoregression, as it includes the lagged value of the dependent variable (consumption) as one of the explanatory variables. The rationale behind this model is that consumers tend not to change their consumption habits quickly due to psychological, technological, or institutional reasons. If the lagged term (previous consumption) is neglected in the equation, the resulting error term will exhibit a systematic pattern. This pattern reflects the influence of lagged consumption on current consumption, leading to autocorrelation in the error term.

- **"Manipulation" of Data:** In empirical analysis, manipulating raw data, such as averaging monthly observations to create quarterly data, can introduce autocorrelation. This smoothing process dampens fluctuations, creating a systematic pattern in the data. Other manipulation techniques, like interpolation or extrapolation, can also impose patterns that might not exist in the original data, leading to autocorrelation. Essentially, data

"massaging" techniques can inadvertently affect the statistical properties of the data, including the presence of autocorrelation.

- **Data Transformation:** When studying relationships between variables like consumption expenditure and income, the level form and the difference form may be transformed into logarithms or percentage changes. This transformation into the growth form can lead to autocorrelation in the error term. Even if the original error term satisfies the standard OLS assumptions, including no autocorrelation, the transformed error term in the dynamic regression models (involving lagged regressands) can become autocorrelated. This illustrates how data transformation can introduce autocorrelation into the analysis.
- **Nonstationarity:** In dealing with time series data, nonstationarity can be a source of autocorrelation. A time series is considered stationary if its characteristics like mean, variance, and covariance do not change over time. If these characteristics do change, the time series is nonstationary. In a regression model, if both dependent and independent variables are nonstationary, the error term may also become nonstationary, leading to autocorrelation. This phenomenon can be either positive or negative, although most economic time series generally exhibit positive autocorrelation. Nonstationarity, therefore, is a key factor that can introduce autocorrelation into time series analysis.

These sources reflect the complexity of time series data and the challenges in modeling relationships between variables. Understanding and addressing these sources is essential for accurate and reliable empirical analysis, as autocorrelation can affect the efficiency and validity of statistical inferences. Whether it's through the nature of the data, the modeling approach, or the transformation techniques, autocorrelation can emerge, requiring careful consideration and correction in econometric practice.

7.4.2. Consequences of Autocorrelation

The consequences of using Ordinary Least Squares (OLS) in the presence of autocorrelation are multifaceted and can lead to significant issues in statistical analysis. Here's a revised summary with academic citations:

- **OLS Estimators Remain Unbiased but Lose Efficiency:** In the presence of autocorrelation, the OLS estimators continue to be unbiased, consistent, and asymptotically normally distributed. However, they lose their efficiency, meaning that they are no longer the Best Linear Unbiased Estimators (BLUE). This inefficiency can lead to incorrect inferences in hypothesis testing.

- **Inaccurate Variance Estimation:** The situation becomes potentially serious if the variance of the OLS estimators is calculated without considering autocorrelation. The residual variance is likely to underestimate the true value, leading to errors in statistical analysis. This underestimation can also result in an overestimation of the R^2 value, which measures the goodness of fit of the model.

- **Inapplicability of Usual Tests:** Due to the loss of efficiency, the usual t, F, and χ^2 tests cannot be legitimately applied. This undermines the reliability of the statistical conclusions drawn from the model (Page 474; Cai et al., 2021).

- **Remedial Measures Depend on the Nature of Autocorrelation:** The appropriate remedy for dealing with autocorrelation depends on the specific nature of the interdependence among the disturbances. Understanding the underlying cause of autocorrelation is essential for applying the correct corrective measures.

- **Potential Misinterpretation of Results:** If autocorrelation is disregarded, and the usual assumptions of the classical model are mistakenly believed to hold true, this can lead to serious misinterpretation of the results. The consequences of this oversight can be far-reaching, affecting both the understanding of the underlying phenomena and the decision-making processes based on the model's results.

In conclusion, the presence of autocorrelation in a regression model poses significant challenges to the use of OLS. It affects the efficiency of the estimators, the accuracy of variance estimation, and the applicability of standard statistical tests. Careful consideration of the nature of autocorrelation and the application of appropriate remedial measures are essential to ensure the validity and reliability of the statistical analysis.

7.4.3. Methods of Detection of Autocorrelation

7.4.3.1. Graphical Method

In the realm of statistical analysis, the assumption of non-autocorrelation is often pivotal, especially in the context of the classical model. However, the population disturbances, denoted by u_t, are not directly observable. Instead, their proxies, the residuals $\hat{u}_t$, are obtained through the Ordinary Least Squares (OLS) procedure. The Graphical Method serves as an instrumental approach to discerning the presence of autocorrelation by examining these residuals.

The method commences by plotting the actual or standardized residuals. This can be further augmented by plotting current residuals against past residuals. Such graphical representations can reveal patterns or systematic behaviors in the residuals, indicative of autocorrelation. For instance, a correlogram, which is a plot of the sample autocorrelation function against the lag, can be used to ascertain if a particular time series is stationary. In the context of a purely white noise process, the autocorrelations at various lags hover around zero, and if the correlogram of an actual time series resembles that of a white noise time series, it can be inferred that the time series is probably stationary.

Furthermore, the examination of residuals is not only a visual diagnostic tool to detect autocorrelation but also serves to identify other issues such as heteroscedasticity or model specification errors. Distinct patterns in the plot of residuals may reveal such errors, thereby underscoring the multifaceted utility of this method.

In conclusion, the Graphical Method is a versatile and intuitive approach to detecting autocorrelation. By visually representing the residuals and their relationships, it provides insights into the underlying structure of the data. This method, although seemingly simple, is imbued with profound implications for statistical modeling and hypothesis testing.

7.4.3.2. Formal Tests for Detection of Autocorrelation

7.4.3.2.1. The Runs Test

Autocorrelation, a phenomenon where error terms in a time series are correlated with each other, can lead to inefficiencies and biases in the estimation process. Detecting autocorrelation is thus a critical step in ensuring the robustness of statistical models. Among the various methods to detect autocorrelation, the Runs Test stands out as a formal approach, sometimes also known as the Geary test, a nonparametric test developed by Geary (1970). If there are too many runs, it would mean that in our example the residuals change sign frequently, thus indicating negative serial correlation. Similarly, if there are too few runs, they may suggest positive autocorrelation. The Runs Test generally involves the following steps:

- **Step 1 – Identify Runs:** A run is defined as a sequence of similar observations. In the context of residuals, a run might consist of a sequence of positive or negative values. The test begins by identifying and counting the number of runs in the data.

- **Step 2 – Calculate Expected Number of Runs:** The expected number of runs is calculated based on the assumption that the data is randomly ordered. This involves using the proportions of positive and negative values in the data.
- **Step 3 – Compute the Test Statistic:** The test statistic is calculated using the observed and expected number of runs, along with the variances and standard deviations of these runs.

$$E(R) = \frac{2N_1N_2}{N} + 1 \quad (7.5)$$

$$\sigma_R^2 = \frac{2N_1N_2(2N_1N_2 - N)}{(N)^2(N-1)} \quad (7.6)$$

Where N_1 is total number of + symbols, N_2 is total number of – symbols, N is total number of observations, and R is number of runs.

- **Step 4 – Compare with Critical Value:** The test statistic is compared with a critical value from a standard normal distribution. If the test statistic falls in the critical region, the null hypothesis of no autocorrelation is rejected.
- **Step 5 – Interpret the Result:** If the null hypothesis is rejected, it indicates the presence of autocorrelation in the data.

The Runs Test is a versatile tool not only for detecting autocorrelation but also for identifying other underlying patterns or systematic behaviors in the data. It's worth noting that the Runs Test is one among several tests for autocorrelation, each with its unique characteristics and applications.

7.4.3.2.2. Durbin-Watson *d* Test

The Durbin-Watson d test is a renowned statistical method used to detect the presence of autocorrelation in the residuals from a regression analysis.

- **Step 1 – Estimate the model:** Regress the original model and obtain the residuals $\hat{u}_t$.
- **Step 1 – Durbin-Watson *d* statistic:** he Durbin-Watson d statistic is defined as:

$$d = \frac{\sum_{t=2}^{t=n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n}\hat{u}_t^2} = \frac{\sum \hat{u}_t^2 + \sum \hat{u}_{t-1}^2 - 2\sum \hat{u}_t\hat{u}_{t-1}}{\sum \hat{u}_t^2} \quad (7.7)$$

Since $\sum \hat{u}_t^2$ and $\sum \hat{u}_{t-1}^2$ differ in only one observation, they are approximately equal. Let use define $\hat{\rho} = \frac{\sum \hat{u}_t\hat{u}_{t-1}}{\sum \hat{u}_t^2}$, therefore,

$$d \approx 2\left(1 - \frac{\sum \hat{u}_t\hat{u}_{t-1}}{\sum \hat{u}_t^2}\right) = 2(1 - \hat{\rho}) \quad (7.8)$$

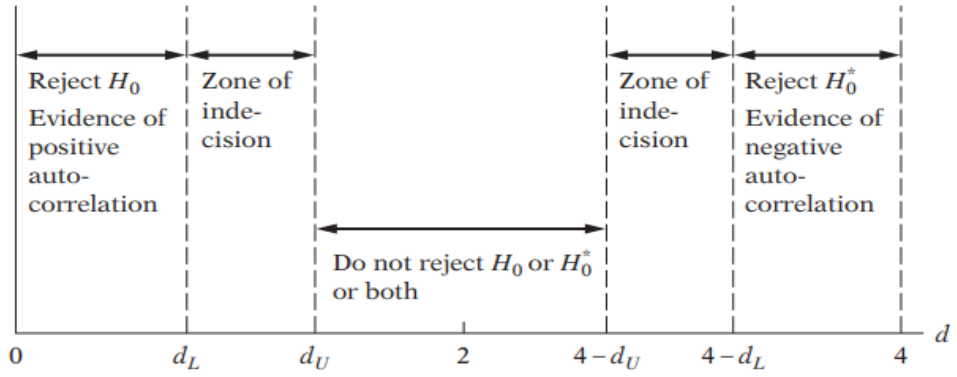
Since $-1 \leq \rho \leq +1$, this implies that d lies within the bounds of 0 to 4:

$$0 \leq d \leq +4 \quad (7.9)$$

- **Step 3 – Find Critical Values:** For the given sample size and given number of explanatory variables, find out the critical d_L and d_U values.
- **Step 4 – Decision Rules:** The value of *d* is then compared with tabulated values, under the null hypothesis $H_0: = \rho = 0$, to make a decision regarding autocorrelation:
    - No positive autocorrelation: Reject if $0 < d < d_L$
    - No positive autocorrelation: No decision if $d_L \leq d \leq d_U$
    - No negative correlation: Reject if $4 - d_L < d < 4$
    - No negative correlation: No decision if $4 - d_U \leq d \leq 4 - d_L$
    - No autocorrelation, positive or negative: Do not reject if $d_U < d < 4 - d_U$

  The information about decision rules is shown in the following Figure 7.1:

  Figure 7.1: Durbin-Watson *d* Statistic



- **Step 5 – Interpretation:** The value of *d* provides evidence regarding the presence or absence of positive or negative serial correlation in the residuals.

The Durbin-Watson *d* test, despite its hoary past, has both merits and limitations. On the one hand, it's a popular and routinely used test for detecting serial correlation, especially in economic models involving time series data. On the other hand, the test has severe limitations, particularly when the value falls in the indecisive zone, leading to inconclusive evidence regarding autocorrelation. Moreover, the test's validity can be compromised in cases where the assumptions underlying the model are not met, such as when the explanatory variables are nonstochastic. Some authors have even contended that the Durbin-Watson statistic may not be useful in econometrics involving time series data, suggesting more useful tests based on large samples.

In academic literature, the Durbin-Watson *d* test has been extensively discussed and critiqued. For instance, Savin and White (1977) extended the original Durbin-Watson table, providing insights into its application with extremely small samples or many regressors. Additionally, studies by Wooldridge (2002) have contributed to the understanding of autocorrelation and the limitations of the Durbin-Watson *d* test in various contexts.

7.4.3.2.3. Breusch–Godfrey Test

The Breusch-Godfrey (BG) test, also known as the Lagrange multiplier test, is a general test for detecting autocorrelation in a time series. The Breusch-Godfrey test is an advanced method for detecting autocorrelation that transcends some of the limitations of other tests. It is general in the sense that it allows for nonstochastic regressors, such as the lagged values of the regressand, higher-order autoregressive schemes like AR(1), AR(2), etc., and simple or higher-order moving averages of white noise error terms. Following are the steps involved:

- **Step 1 – Estimate the model:** Start by estimating the regression model that you want to test for autocorrelation and obtain the residuals $\hat{u}_t$.
- **Step 2 – Estimate the model:** Regress $\hat{u}_t$ on the original X_t (if there is more than one X variable in the original model, include them also) and $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-p}$, where the latter are the lagged values of the estimated residuals in step 1. Thus, if $p = 4$, we will introduce four lagged values of the residuals as additional regressors in the model. Note that to run this regression we will have only $(n - p)$ observations. In short, run the following regression:

  $$\hat{u}_t = \alpha_1 + \alpha_2 X_t + \hat{\rho}_1 \hat{u}_{t-1} + \hat{\rho}_2 \hat{u}_{t-2} + \cdots + \hat{\rho}_4 \hat{u}_{t-4} + \varepsilon_t \qquad (7.10)$$

  and obtain R^2 from this auxiliary regression.
- **Step 3 – Calculate χ^2:** If the sample size is large (technically, infinite), Breusch and Godfrey have shown that $(n - p)R^2 \sim \chi_p^2$, that is, asymptotically, $n - p$ times the R^2 value obtained from the auxiliary regression follows the χ^2 distribution with p df.
- **Step 4 – Interpretation:** If in an application, calculated χ^2 exceeds the critical chi-square value at the chosen level of significance, we reject the null hypothesis (H_0: $\rho_1 = \rho_2 = \cdots = \rho_p = 0$), in which case at least one ρ in is statistically significantly different from zero.

The following practical points about the BG test may be noted:

- The regressors included in the regression model may contain lagged values

of the regressand Y, that is, Y_{t-1}, Y_{t-2}, etc., may appear as explanatory variables. Contrast this model with the Durbin–Watson test restriction that there may be no lagged values of the regressand among the regressors.

- BG test is applicable even if the disturbances follow a p^{th} order moving average (MA) process, that is, the u_t are generated as follows:

$$u_t = \varepsilon_t + \lambda_1 \varepsilon_{t-1} + \lambda_2 \varepsilon_{t-2} + \cdots + +\lambda_p \varepsilon_{t-p} \quad (7.11)$$

  Where ε_t is a white noise error term, that is, it satisfies all the classical assumptions.

- If $\rho = 1$, meaning first order autoregression, then the BG test is known as Durbin's M test.

The Breusch-Godfrey test, renowned for its generality, offers significant advantages over other autocorrelation detection methods. It accommodates both autoregressive (AR) and moving average (MA) error structures, as well as the presence of lagged values, making it a versatile tool in the field of econometrics. Moreover, its statistical power in both large and small samples sets it apart from other tests, enhancing its applicability across various research contexts. However, this test is not without its challenges. Its mathematical complexity and intricate reasoning behind the test statistic may pose difficulties for those less versed in advanced statistical methods. Additionally, being a large-sample test, its application in small samples is not strictly justified, which may limit its utility in certain scenarios. Thus, while the Breusch-Godfrey test's robustness and flexibility make it a preferred choice for many researchers, its complexity and large-sample orientation may present obstacles in specific applications.

The Breusch-Godfrey test's ability to handle nonstochastic regressors and higher-order schemes makes it a valuable tool in the econometrician's toolkit. Its general nature and statistical power in both large and small samples contribute to its preference over other tests, despite some complexities and limitations.

In the multifaceted realm of econometrics, the detection of autocorrelation is a pivotal task, and various tests have been developed to address this challenge. The Durbin-Watson *d* test, with its simplicity and routine application, offers a quick assessment but may fall short in complex scenarios. The Breusch-Godfrey test, on the other hand, provides a more nuanced and general approach, accommodating various statistical complexities, but at the cost of mathematical intricacy. The Runs Test offers versatility but may lack the specificity required in certain contexts. Each test, with its unique characteristics, advantages, and limitations, serves needs and research contexts. The choice of a specific test often hinges on the underlying data structure, model assumptions, and the researcher's specific requirements. In

summary, the landscape of autocorrelation detection is rich and varied, and the judicious selection of a test requires a careful consideration of the research question, data properties, and the trade-offs between simplicity, generality, and statistical power.

7.4.4. Remedial Measures of Autocorrelation

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.

7.4.4.1. When ρ is Known

If the coefficient of first-order autocorrelation is known, the problem of autocorrelation can be easily solved. As a starter, consider a two-variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{7.12}$$

Now assume that the error term follows a AR(1) scheme, namely:

$$u_t = \rho u_{t-1} + \varepsilon_t \tag{7.13}$$

If the model holds true at time t, it also holds true at time $t-1$. Hence,

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \tag{7.14}$$

Multiplying the equation (7.14) by ρ:

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \tag{7.15}$$

Subtracting equation (7.15) from (7.12)

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t \tag{7.16}$$

Where $\varepsilon_t = u_t - \rho u_{t-1}$, we can express equation (7.16) as:

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t \tag{7.17}$$

Where $Y_t^* = (Y_t - \rho Y_{t-1})$, $X_t^* = (X_t - \rho X_{t-1})$, $\beta_1^* = \beta_1(1-\rho)$ and $\beta_2^* = \beta_2$. Since the error term ε_t 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.

7.4.4.2. When ρ is Unknown

Because ρ is rarely known in practice, we cannot apply procedure mentioned in previous section. However, there are number of possibilities to solve the problem of autocorrelation.

7.4.4.2.1 The First Difference Method

Since ρ lies between 0 and ±1, one could start from two extreme positions. If $\rho = +1$, the generalized difference equation (7.16) reduces to the first difference equation:

$$Y_t - Y_{t-1} = \beta_2(X_t - X_{t-1}) + (u_t - u_{t-1}) \quad (7.18)$$

$$\triangle Y_t = \beta_2 \triangle X_t + \triangle u_t \quad (7.19)$$

where $\triangle$ is the first difference operator. Since the error term in Eq. (7.19) is free from (first-order) serial correlation, we can run the regression using OLS. The first-difference transformation may be appropriate if the coefficient of autocorrelation is very high, say in excess of 0.8, or the Durbin–Watson d is quite low. Maddala has proposed this rough rule of thumb: Use the first-difference form whenever $d < R^2$. An interesting feature of the first-difference model is that there is no intercept in it.

7.4.4.2.2. ρ Based on Durbin-Watson d Statistics

If the first-difference transformation is not applicable due to ρ not being sufficiently close to unity, we can resort to a straightforward method to estimate ρ. This estimation can be derived from the previously established relationship between d and ρ as expressed in Eq. (7.9), allowing us to calculate the value of accordingly.

$$\hat{\rho} \approx 1 - \frac{d}{2} \quad (7.20)$$

In reasonably large samples, one can extract the value of ρ from Equation (7.20) and utilize it to transform the data, as demonstrated in the generalized difference equation (7.16). It's important to recognize, however, that the relationship between ρ and d as given in Equation (7.20) might not be applicable or hold true in the context of small samples.

7.4.4.2.3. ρ Estimated from the Residuals

If the AR(1) scheme $u_t = \rho u_{t-1} + \varepsilon_t$ is valid, a straightforward method to estimate ρ is to perform a regression of the residuals $\hat{u}_t$, for the $\hat{u}_t$ are consistent estimators of the true u_t, as previously observed. We run the following regression:

$$\hat{u}_t = \rho \hat{u}_{t-1} + v_t \quad (7.21)$$

Where v_t is the error term of this regression. Note that as OLS residuals sum to zero, there is no need to introduce the intercept term.

7.4.4.2.4. Iterative Methods of Estimating ρ

The iterative methods of estimating ρ provide a nuanced approach to understanding autocorrelation, going beyond single estimates to offer successive approximations. These methods begin with an initial value of ρ and refine it through iterations. Among the various iterative methods, some notable ones include the Cochrane–Orcutt iterative procedure, the Cochrane–Orcutt two-step procedure, the Durbin two-step procedure, and the Hildreth–Lu scanning or search procedure.

The most popular among these is the Cochrane–Orcutt iterative method. Unlike other methods that provide only a single estimate, iterative methods work by successive approximation, continually refining the estimate of ρ. This iterative approach allows for more flexibility and can be used to estimate not only an AR(1) scheme but also higher-order autoregressive schemes, such as an AR(2) model represented by $\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + v_t$. Once the two ρ values are obtained, one can easily extend the generalized difference equation (7.17).

The ultimate objective of these iterative methods is to provide an estimate of ρ that can be used to obtain Generalized Least Squares (GLS) estimates of the parameters. One distinct advantage of the Cochrane–Orcutt iterative method is its ability to handle complex autoregressive schemes, making it a versatile tool in econometric analysis. With the advent of modern computing, these iterative methods can now be efficiently implemented, further enhancing their applicability and utility in empirical research.

7.5. Self-Assessment Questions

- What is autocorrelation, and how does it manifest in time-series data?
- How can autocorrelation impact statistical inference and model interpretation? Provide examples.
- Explain how the graphical method can be used to detect autocorrelation? What are its limitations?
- Compare and contrast the Runs Test, Durbin-Watson d test, and Breusch–Godfrey test. When might you choose one over the others?
- Describe the steps involved in remedying autocorrelation when ρ is known and when ρ is unknown? Provide examples of each.
- Explain the First Difference Method and its application in addressing autocorrelation.
- Discuss the Cochrane–Orcutt iterative method and its significance in estimating ρ.
- Given a dataset with autocorrelation, outline the steps you would take to detect and remedy the issue.
- What are the advantages and disadvantages of different tests for detecting autocorrelation? Provide examples.
- Reflect on the importance of understanding and addressing autocorrelation in data analysis. How might ignoring autocorrelation lead to incorrect conclusions?
- Describe how modern computing tools can be leveraged to detect and remedy autocorrelation. Provide examples if possible.
- What further reading or exploration might you undertake to deepen your understanding of autocorrelation and related statistical concepts?

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Cai, C., Li, G., Chi, Y., Poor, H., & Chen, Y. (2021). Subspace Estimation from Unbalanced and Incomplete Data Matrices: $l_{2,\infty}$Statistical Guarantees. Link.
- Geary, R. C. (1970). Relative Efficiency of Count Sign Changes for Assessing Residual Autoregression in Least Squares Regression. *Biometrika*, 57, pp. 123–127.
- Griffiths, W. E., Hill, R. C., & Judge, G. G. *Learning and Practicing Econometrics*, Wiley.
- Kmenta, J. *Elements of Econometrics*, Latest edition, Macmillan, New York.
- Savin, N. E., & White, K. J. (1977). The Durbin-Watson Test for Serial Correlation with Extreme Small Samples or Many Regressors. *Econometrica*, 45, 1989–1996.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 08

MODEL SPECIFICATION AND DIAGNOSTIC TESTING

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

8.1. INTRODUCTION

In the complex realm of empirical modeling, the process of selecting the most suitable model and understanding its underlying assumptions is paramount. The Unit delves into the multifaceted aspects of model specification, exploring the criteria for selecting models that best represent the underlying data and theory. It sheds light on the different types of specification errors that can occur and the potential consequences these errors may have on the validity and reliability of the model. The Unit also examines various tests to detect and rectify these errors, emphasizing the importance of accurate measurements. A nuanced discussion is presented on the comparison between nested and non-nested models, elucidating the methods to test non-nested hypotheses. Furthermore, the Unit explores the intricate balance between model complexity and fit, providing insights into the criteria used in both nested and non-nested models. Overall, the Unit offers a comprehensive guide to understanding, selecting, and validating empirical models, equipping readers with the tools to navigate the complexities of econometric analysis.

8.2. OBJECTIVES

After reading the unit, the students will be able to:

- **understand Model Selection Criteria:** Grasp the fundamental principles and methods used in selecting appropriate models, including understanding the difference between in-sample and out-of-sample forecasting.
- **identify Types of Specification Errors:** Learn to recognize various specification errors such as omission of relevant variables, incorrect functional forms, and errors in measurement.
- **analyze Consequences of Model Specification Errors:** Evaluate the impact of specification errors on the model, including understanding the concepts of underfitting and overfitting.
- **apply Tests of Specification Errors:** Master various tests like Durbin-Watson d statistic, Ramsey's RESET test, and Lagrange Multiplier test to detect and correct specification errors.
- **assess Errors of Measurements:** Understand the nuances of errors in the measurement of dependent and explanatory variables and their implications on the model.
- **distinguish Between Nested and Non-Nested Models:** Comprehend the differences between nested and non-nested models and their applications in different contexts.

- **conduct Tests of Non-Nested Hypotheses:** Learn to apply specific tests like the J test to evaluate non-nested hypotheses.
- **evaluate Model Selection Criteria in Nested and Non-Nested Models:** Understand and apply various criteria such as AIC, SIC, and Mallows's Cp criterion for model selection in both nested and non-nested contexts.
- **develop Critical Thinking and Analytical Skills:** Cultivate the ability to critically analyze models, recognize potential flaws, and apply appropriate remedies.
- **engage in Practical Application:** Apply the theoretical knowledge gained to real-world data and scenarios, enhancing practical skills in model selection and validation.

By achieving these objectives, students will be well-equipped to navigate the complexities of model specification and selection, fostering a robust understanding of econometric analysis and its practical applications.

8.3. Major Topics

- Model Selection Criteria
- Types of Specification Errors
- Consequences of Model Specification Errors
- Tests of Specification Errors
- Errors of Measurements
- Nested Versus Non-Nested Models
- Tests of Non-Nested Hypothesis
- Model Selection Criteria in Nested and Non-Nested Models

8.4. Summary of the Units

8.4.1. Model Selection Criteria

In the context of empirical analysis, the selection of an appropriate model is guided by several vital criteria, as delineated by Hendry and Richard (1983):

- **Data Admissibility:** The model must produce predictions that are logically feasible. In other words, the outcomes derived from the model must be within the realm of possibility.
- **Consistency with Theory:** The model should align with sound economic principles. For instance, if Milton Friedman's permanent income hypothesis is valid, the intercept value in the regression of permanent consumption on permanent income would be anticipated to be zero.
- **Weakly Exogenous Regressors:** The explanatory variables or regressors in the model should not be correlated with the error term. In certain scenarios, the exogenous regressors may even be strictly exogenous, meaning they are independent of the current, future, and past values of the error term.
- **Parameter Constancy:** The values of the parameters within the model must remain stable. If they fluctuate, forecasting becomes challenging. As Friedman insightfully observed, the validity of a hypothesis or model is best tested by comparing its predictions with real-world outcomes. Without parameter constancy, such predictions lack reliability.
- **Data Coherency:** The residuals estimated from the model must exhibit pure randomness, or technically, white noise. If the regression model is well-specified, the residuals must be white noise. A deviation from this pattern

indicates a specification error in the model, signaling a need for further investigation.

- **Encompassing Nature:** The chosen model should be comprehensive, encompassing or including all rival models. It should be capable of explaining the results of other models, ensuring that no other models can provide an improvement over the selected one.

These criteria collectively form a robust framework for model selection, ensuring that the chosen model is not only theoretically sound but also empirically reliable. They guide the researcher in balancing the complexities of economic theory with the practicalities of data analysis, leading to models that provide meaningful insights and accurate predictions.

In the process of developing an empirical model, researchers may encounter various challenges that lead to specification errors. These errors can be broadly categorized into the following types:

- **Omission of Relevant Variables:** Leaving out variables that have a significant impact on the dependent variable can lead to biased and inconsistent estimates.
- **Inclusion of Unnecessary Variables:** Conversely, including variables that do not have a significant relationship with the dependent variable can lead to inefficiency and overfitting.
- **Adoption of the Wrong Functional Form:** Choosing an incorrect functional relationship between the dependent and independent variables can lead to systematic errors in the estimated parameters.
- **Errors of Measurement:** Inaccuracies in the data collection process can introduce noise and bias into the model, affecting the reliability of the results.
- **Incorrect Specification of the Stochastic Error Term:** Misrepresenting the error term can lead to violations of the assumptions underlying the statistical techniques used, affecting the validity of the inferences drawn.
- **Assumption that the Error Term is Normally Distributed:** This assumption may not always hold true, and its violation can affect the properties of the estimators used.

It is also essential to recognize the difference between model specification errors and model mis-specification errors. The former refers to situations where there is a

"true" model in mind, but somehow the correct model is not estimated. This includes the first four types of errors mentioned above.

On the other hand, model mis-specification errors occur when the true model is unknown from the outset. This can lead to competing models, each with its own underlying assumptions and focus. For example, the controversy between Keynesians and monetarists illustrates this type of error. While monetarists emphasize the role of money in explaining changes in GDP, Keynesians focus on government expenditure. These represent two distinct and competing models, reflecting differing economic theories.

In summary, understanding and addressing specification errors is crucial in empirical modeling. These errors can significantly impact the validity and interpretability of the model, and recognizing the nuances between different types of errors helps in designing more robust and accurate empirical analyses.

8.4.2. Consequences of Model Specification Errors

The consequences of model specification errors are critical to understand, particularly in the context of underfitting and overfitting a model.

8.4.2.1. Underfitting a Model (Omitting Relevant Variables):

- **Bias in Estimation:** Omitting a relevant variable that is correlated with the included explanatory variable leads to biased estimators. This bias affects not only the omitted variable but also the estimators of other parameters.
- **Inefficiency:** The OLS estimators are no longer BLUE (Best Linear Unbiased Estimators), leading to inefficiency in the estimation.
- **Inconsistency in Hypothesis Testing:** The t and F tests may no longer be valid, leading to incorrect conclusions about the significance of variables.

8.4.2.2. Overfitting a Model (Including Unnecessary Variables):

- **Loss of Efficiency:** Including unnecessary variables leads to a loss in the efficiency of the estimators and may also lead to the problem of multicollinearity.
- **Loss of Degrees of Freedom:** Overfitting consumes degrees of freedom, reducing the power of statistical tests.
- **Complexity and Misinterpretation:** An overfitted model may capture noise rather than the underlying relationship, leading to misinterpretation of the results.

The balance between underfitting and overfitting is delicate. Underfitting ignores essential complexities in the data, leading to a model that is too simple to capture the underlying relationships. Overfitting, on the other hand, includes unnecessary complexities, capturing random noise in the data and leading to a model that does not generalize well to new data.

In general, the best approach is to include only explanatory variables that, on theoretical grounds, directly influence the dependent variable and that are not accounted for by other included variables. Knowing the consequences of specification errors is essential, but finding out whether one has committed such errors is quite another challenge. Specification biases often arise inadvertently, perhaps from our lack of understanding of the underlying economic theory or from data limitations.

The practical question is not why specification errors are made, for they generally are, but how to detect them. Once it is found that specification errors have been made, the remedies often suggest themselves. If a variable is inappropriately omitted from a model, the obvious remedy is to include that variable in the analysis, assuming the data on that variable are available.

In conclusion, model specification errors, whether through underfitting or overfitting, have significant consequences in econometric analysis. They can lead to biased and inefficient estimators, incorrect hypothesis testing, and misleading interpretations. Careful consideration of theory, data, and statistical diagnostics is essential to avoid these pitfalls and to develop models that provide reliable insights into the phenomena under study.

8.4.3. Tests of Specification Errors

In empirical testing, the certainty that the chosen model is entirely accurate is often elusive. Researchers develop models based on theory, introspection, and previous empirical work, and then subject them to testing. The adequacy of the model is determined post-analysis by examining various features such as the R^2 value, estimated t ratios, signs of the estimated coefficients, and the Durbin–Watson statistic. If these diagnostics align well, the model is considered a fair representation of reality. Conversely, if the results are not satisfactory, concerns about model adequacy arise, leading to a search for potential errors such as omitted variables, incorrect functional form, or issues related to serial correlation. Here's a detailed look at the tests of specification errors, including various methods and their implications:

8.4.3.1. Examination of Residuals

Residuals are the differences between observed values and predicted values. By examining the residuals, one can detect patterns that may indicate specification errors. If the residuals exhibit noticeable patterns, it indicates that there might be some specification error in the model.

8.4.3.2. Durbin-Watson *d* Statistic

The Durbin-Watson d statistic is used to detect first-order serial correlation in the error terms. It is a popular method but has limitations, especially if it falls in the indecisive zone, where one cannot conclude whether or not there is autocorrelation. To employ the Durbin–Watson test for detecting model specification errors, the following steps are undertaken:

- **Step 1 – Calculate Residuals:** From the assumed model, the ordinary least squares (OLS) residuals are obtained.
- **Step 2 – Order Residuals:** If the model is suspected to be mis-specified due to the exclusion of a relevant explanatory variable, such as Z, the residuals from Step 1 are ordered according to increasing values of Z. The variable Z could be one of the X variables included in the assumed model or a function of that variable, such as X^2 or X^3.
- **Step 3 – Compute the d Statistic:** The d statistic is computed from the ordered residuals using the formula:

  $$d = \frac{\sum_{t=2}^{n}(\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n}\hat{u}_t^2} \quad (8.1)$$

  Here, the subscript t is the index of observation and does not necessarily imply that the data are time series.
- **Step 3 – Interpret the d Value:** By referring to the Durbin–Watson tables, if the estimated d value is significant, the hypothesis of model mis-specification can be accepted. If this is the case, appropriate remedial measures will naturally suggest themselves.

This process allows for a systematic examination of the model to identify potential specification errors, providing a pathway to refine the model for more accurate representation of the underlying phenomena.

8.4.3.3. Ramsey's RESET Test

Ramsey's RESET (regression specification error test) is a general test for specification error. It can detect a range of alternatives and indicate something

wrong under the null hypothesis without necessarily giving clear guidance as to what alternative hypothesis is appropriate. The Ramsey's RESET is a general test for model specification errors. Here's how it can be applied:

- **Step 1 – Calculate Estimated Values:** From the chosen model, obtain the estimated values of Y, denoted as $\hat{Y}_i$.
- **Step 2 –** Introduce Additional Regressors: Rerun the model by introducing $\hat{Y}_i$ in some form as an additional regressor(s). If there is a curvilinear relationship between the residuals $\hat{u}_i$ and $\hat{Y}_i$, one can introduce $\hat{Y}_i^2$ and $\hat{Y}_i^3$ as additional regressors. Thus, the model becomes:
$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i \quad (8.2)$$
- **Step 3 – Compute F Statistic:** Let R^2_{new} be the R^2 obtained from the new model, and R^2_{old} be that obtained from the original model. The F test can be used to determine if the increase in R^2 is statistically significant:
$$F = \frac{(R^2_{new} - R^2_{old})/number\ of\ new\ regressors}{(1 - R^2_{new})/(n - number\ of\ parameters\ in\ the\ new\ model)} \quad (8.3)$$
- **Step 4 – Interpret the F Value:** If the computed F value is significant, such as at the 5 percent level, one can accept the hypothesis that the original model is mis-specified.

The RESET test is a powerful tool in detecting model specification errors, particularly when there is a concern that some nonlinear combination of the fitted values might be correlated with the residuals. By introducing these nonlinear combinations as additional regressors, the test can provide insights into whether the functional form of the model is correctly specified.

8.4.3.4. Lagrange Multiplier (LM) Test for Adding Variables

This test is an alternative to Ramsey's RESET test and is used to determine if additional variables should be included in the model. It helps in identifying if the model is a restricted version of another and whether adding variables would improve the specification. The Lagrange Multiplier (LM) test is a statistical test used to determine whether a restricted model is appropriate. Here's how the LM test can be applied to detect specification errors:

- **Step 1 – Estimate the Restricted Regression:** Estimate the restricted regression $Y_i = \lambda_1 + \lambda_2 X_i + u_{1i}$ by Ordinary Least Squares (OLS) and obtain the residuals, denoted as $\hat{u}_i$.
- **Step 2 – Consider Unrestricted Regression:** If the unrestricted regression $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i$ is the true regression, the residuals

obtained should be related to the squared and cubed output terms, such as X_i^2 and X_i^3.

- **Step 3 – Regress Residuals:** Regress the $\hat{u}_i$ obtained in Step 1 on all the regressors (including those in the restricted regression). In this case, the regression becomes:

$$\hat{u}_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 X_i^3 + v_i \quad (8.4)$$

where v is an error term with the usual properties.

- **Step 4 – Compute Chi-Square Statistic:** For large sample sizes, it has been shown that n (the sample size) times the R^2 estimated from the auxiliary regression follows the chi-square distribution with degrees of freedom equal to the number of restrictions imposed by the restricted regression. Symbolically:

$$nR^2 \sim \chi^2 \quad (8.5)$$

- **Step 4 – Interpret the Chi-Square Value:** If the chi-square value obtained exceeds the critical chi-square value at the chosen level of significance, the restricted regression is rejected. Otherwise, it is not rejected.

The LM test is a powerful tool for model specification, allowing researchers to test whether certain restrictions are valid. By comparing the restricted and unrestricted models, it provides insights into whether the simpler model is sufficient or if additional complexity is warranted. It's particularly useful in large samples where asymptotic properties come into play.

In summary, the selection of an appropriate model and the detection of specification errors are vital steps in empirical analysis. Various tests and methods are available to ensure that the chosen model is a fair representation of reality and to detect and correct any specification or mis-specification errors. These methods provide a robust framework for understanding and interpreting the data, thereby enhancing the reliability and validity of the findings. The integration of these tests with theoretical understanding and computational tools can lead to more precise and insightful empirical studies.

8.4.4. Errors of Measurement

Errors of measurement in econometric modeling can have significant implications for the accuracy and reliability of the results. These errors can occur in both the dependent variable (Y) and the explanatory variable (X), and each type of error has distinct consequences.

- **Errors of Measurement in the Dependent Variable Y:** When errors of measurement occur in the dependent variable Y, they do not necessarily

destroy the unbiasedness property of the Ordinary Least Squares (OLS) estimators. Consider the model $Y_i^* = \alpha + \beta X_i + u_i$, where Y_i^* is not directly measurable, and we use an observable expenditure variable Y_i instead. The errors of measurement in Y do not affect the unbiasedness of the estimators, but they do impact the variances and standard errors of β. Specifically, the estimated variances are larger than in the case where there are no such errors of measurement.

Although the errors of measurement in the dependent variable still give unbiased estimates of the parameters, the estimated variances are now larger. This does not necessarily invalidate the model but may reduce the precision of the estimates.

- **Errors of Measurement in the Explanatory Variable X:** Errors of measurement in the explanatory variable X pose a more serious problem. Consider the model $Y_i = \alpha + \beta X_i^* + u_i$, where X_i^* represents permanent income. If there are measurement errors in X, the estimators become inconsistent, making consistent estimation of the parameters impossible. If β is assumed positive, β will underestimate β, meaning it is biased toward zero. If there are no measurement errors in X, β will provide a consistent estimator of β.

  Measurement errors in the explanatory variable make consistent estimation of the parameters impossible. This can lead to biased estimates and seriously undermine the validity of the model.

Addressing errors of measurement is not straightforward. The use of instrumental or proxy variables is theoretically attractive but not always practical. It is crucial to measure the data as accurately as possible, and researchers should be careful in stating the sources of their data, how they were collected, and what definitions were used.

In conclusion, errors of measurement are a potentially troublesome problem in econometric modeling, constituting an example of specification bias. They can lead to underestimation or overestimation of parameters, affecting the model's adequacy and reliability. The consequences of these errors underscore the importance of accurate data collection and careful consideration of the underlying assumptions in the modeling process. The presence of errors in measurement in the regressors can lead to biased as well as inconsistent OLS estimators, making the remedies often not easy and emphasizing the importance of careful data collection.

8.4.5. Nested vs Non-Nested Models

In the realm of econometric modeling, the concept of nested and non-nested models plays a pivotal role in understanding and testing the specification of models. These models are instrumental in determining the relationships between variables and can be used to test different hypotheses or theories.

- **Nested Models:** Nested models occur when one model can be derived as a special case of another by imposing certain restrictions on the parameters. Consider two models:

  Model A: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$ (8.6)

  Model B: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$ (8.7)

  Model B is nested in Model A if $\beta_4 = \beta_5 = 0$. If this hypothesis is not rejected, Model A reduces to Model B.

- **Non-Nested Models:** Non-nested models are those where one model cannot be derived as a special case of the other. Consider the following models:

  Model C: $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$ (8.8)

  Model D: $Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + u_i$ (8.9)

  These models are non-nested because they contain different variables (X's and Z's), and one cannot be derived from the other. Even if the same variables are present, different functional forms can make models non-nested, such as:

  Model E: $Y_i = \beta_1 + \beta_2 \ln Z_{2i} + \beta_3 \ln Z_{3i} + u_i$ (8.10)

  Models D and E are non-nested due to the logarithmic transformation of variables.

The distinction between nested and non-nested models is vital in econometric analysis. Nested models allow for straightforward testing using traditional methods, while non-nested models require more complex techniques. Nested models involve a hierarchical relationship where one model can be derived from another by imposing restrictions on the parameters. Non-nested models represent different theories or functional forms and cannot be reduced to one another. Understanding these concepts is essential for selecting the appropriate testing procedures and interpreting the results, reflecting the underlying theories and assumptions that guide empirical analysis.

8.4.6. Tests of Non-Nested Hypothesis

According to Harvey (1990), there are two approaches to testing non-nested hypotheses.

8.4.6.1. The Discrimination Approach

When faced with multiple models, such as Models C and D from previous section, that involve the same dependent variable, a selection must be made based on a goodness-of-fit criterion. Commonly used criteria include the coefficient of determination (R^2) or the adjusted R^2, both of which provide insight into how well the model fits the observed data. However, it's essential to ensure that the dependent variable, or regressand, is consistent across the models being compared.

Beyond these standard measures, other sophisticated criteria are often employed in model selection. These include Akaike's Information Criterion (AIC), Schwarz's Information Criterion (SIC), and Mallows's C_p criterion, each offering a unique perspective on model fit and complexity. Many modern statistical software packages incorporate these criteria into their regression routines, facilitating the comparison process.

Ultimately, the chosen model is typically the one that maximizes the adjusted R^2 or minimizes the values of AIC or SIC. This selection process is vital in empirical analysis, guiding researchers to the model that best represents the underlying data and theoretical considerations.

8.4.6.2. The Discerning Approach

8.4.6.2.1. The Non-Nested F-test or Encompassing F-test

When considering Models C and D, as introduced in equation (8.8) and (8.9) the question arises: how to choose between these two non-nested models? A common approach is to estimate a nested or hybrid model, such as Model F, which encompasses both Models C and D.

$$\text{Model F: } Y_i = \lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \lambda_4 Z_{2i} + \lambda_5 Z_{3i} + u_i \qquad (8.11)$$

In this model, if Model C is correct, certain coefficients will equal zero, and if Model D is correct, other coefficients will equal zero. This can be tested using the non-nested F test.

However, this testing procedure is not without problems. First, if the variables in Models C and D are highly correlated, it may lead to multicollinearity, making it difficult to determine the statistical significance of individual coefficients. In such

a scenario, neither Model C nor Model D may be definitively chosen as the correct model.

Second, the choice of the reference hypothesis or model can influence the outcome. If Model C is chosen as the reference and found to be significant, it may be selected as the correct model. Conversely, if Model D is chosen as the reference and found to be significant, it may be selected instead. This ambiguity in selection, especially in the presence of severe multicollinearity, highlights the complexity of choosing between non-nested models.

Lastly, the artificially nested model F may lack economic meaning, further complicating the decision-making process. The challenges presented by this approach underscore the importance of careful consideration and robust testing when selecting between competing models in empirical analysis.

8.4.6.2.2. Davidson-MacKinnon *J* Test

The J test is used to compare two non-nested models, such as Model C and Model D. The procedure is designed to overcome some of the problems associated with non-nested F testing and is carried out as follows:

- Estimate Model D: Start by estimating Model D and obtaining the estimated Y values, denoted as $\hat{Y}_{D_i}$.

- Add Predicted Y Value to Model C: The predicted Y value from Step 1 is added as an additional regressor to Model C, and the following model is estimated:

  $$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 \hat{Y}_{D_i} + u_i \quad (8.12)$$

  Where $\hat{Y}_{D_i}$ values are obtained from Step 1.

- Test Hypothesis for α_4: Using the t-test, test the hypothesis that $\alpha_4 = 0$.
- Decision Making for Model C: If the hypothesis that $\alpha_4 = 0$ is not rejected, Model C can be accepted as the true model. If the null hypothesis is rejected, Model C cannot be the true model.

- Reverse Roles of Models C and D: Now, estimate Model C first, use the estimated Y values ($\hat{Y}_{C_i}$) from this model as the regressor in model D, and repeat Step 4. More specifically, estimate the following model:

  $$Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 \hat{Y}_{C_i} + u_i \quad (8.13)$$

  Where $\hat{Y}_{C_i}$ are the estimated Y values from Model C. Test the hypothesis that $\beta_4 = 0$. If this hypothesis is not rejected, choose Model D over C. If the hypothesis that $\beta_4 = 0$ is rejected, choose C over D.

The J test is intuitively appealing but has some problems. Since the tests given in Eqs. (8.12) and (8.13) are performed independently, there may be likely outcomes that could lead to ambiguity in the decision-making process. The test's design, which encompasses the principles of the Hendry methodology, aims to determine whether one model contains additional information that would improve the performance of the other, thus guiding the selection between non-nested models.

Testing non-nested models is a multifaceted task that requires careful consideration of the underlying theories and data structures. The Discrimination Approach focuses on selecting the best-fitting model, while the Discerning Approach aims to identify the true model. Various statistical tests, such as the J-Test, Cox Test, and Vuong's Test, provide robust methodologies for comparing non-nested models, each with its unique strengths and applications. These tests play a crucial role in empirical analysis, guiding researchers in model selection and interpretation.

8.4.7. Model Selection Criteria in Nested and Non-Nested Models

In this section, the focus is on the various criteria used to select and compare different models, particularly for forecasting. The discussion differentiates between two types of forecasting: in-sample and out-of-sample. In-sample forecasting evaluates how well the selected model fits the data within the existing sample. Conversely, out-of-sample forecasting examines how the fitted model predicts future values of the dependent variable, considering the values of the independent variables.

8.4.7.1. The R^2 Criterion

The R^2 value is a well-known measure of the goodness of fit in a regression model, defined as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \tag{8.14}$$

However, this metric has several inherent limitations. First, it measures the in-sample goodness of fit, showing how closely the estimated values of the dependent variable match the actual values within the sample. This means that it does not necessarily guarantee good forecasting for out-of-sample observations. Second, when comparing two or more R^2 values, the dependent variable must be the same, limiting its applicability in some comparative scenarios. Third, since R^2 cannot decrease when more variables are added to the model, there may be a temptation to "maximize the R^2" by simply including more variables. While this may increase R^2, it can also inflate the variance of the forecast error, leading to a model that is overfitted and less generalizable. This highlights the need for careful consideration

and potential supplementary criteria when using R^2 for model selection or evaluation.

8.4.7.2. Adjusted R^2 Criterion

The adjusted R^2, denoted by $\bar{R}^2$, was developed by Henry Theil as a way to penalize the addition of regressors that might artificially increase the R^2 value.

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{8.15}$$

Unlike R^2, the adjusted R^2 takes into account the number of regressors in the model, and it will only increase if the absolute t-value of the added variable is greater than 1. This makes $\bar{R}^2$ a more robust measure for comparative purposes, as it avoids the temptation to overfit the model by adding unnecessary variables. However, it's important to note that the comparison using $\bar{R}^2$ is only valid when the dependent variable, or regressand, is the same across the models being compared.

8.4.7.3. Akaike's Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a measure that imposes a penalty for adding regressors to a model, going beyond the adjusted R^2 in this regard. The AIC is defined as:

$$AIC = e^{2k/n}\frac{\sum \hat{u}_i^2}{n} = e^{2k/n}\frac{RSS}{n} \tag{8.16}$$

or, in its log-transformed form:

$$\ln AIC = \frac{2k}{n} + \ln\left(\frac{RSS}{n}\right) \tag{8.17}$$

where k is the number of regressors (including the intercept), n is the number of observations, and RSS is the residual sum of squares. The AIC imposes a harsher penalty than $\bar{R}^2$ for adding more regressors, and the model with the lowest value of AIC is preferred. One of the advantages of AIC is its applicability to both in-sample and out-of-sample forecasting performance, as well as its usefulness for both nested and non-nested models. It has also been employed to determine the lag length in an autoregressive model ($AR(p)$).

8.4.7.4. Schwarz's Information Criterion (SIC)

The Schwarz Information Criterion (SIC), also known as the Bayesian Information Criterion (BIC), is another measure used to select among competing models, and it's similar in spirit to the AIC. The SIC is defined as:

$$SIC = n^{k/n}\frac{\sum \hat{u}_i^2}{n} = n^{k/n}\frac{RSS}{n} \tag{8.18}$$

or, in its log-transformed form:

$$\ln SIC = \frac{k}{n}\ln n + \ln\left(\frac{RSS}{n}\right) \tag{8.19}$$

where k is the number of regressors (including the intercept), n is the number of observations, and RSS is the residual sum of squares. The penalty factor in SIC is $\frac{k}{n}\ln n$, which imposes a harsher penalty than AIC. As with AIC, the model with the lowest value of SIC is preferred. SIC can be used to compare both in-sample and out-of-sample forecasting performance of a model, making it a versatile criterion for model selection.

8.4.7.5. Mallow's C_p Criterion

Mallows's C_p criterion is a statistical measure used for model selection, particularly when choosing the number of regressors in a model. Suppose a model consists of k regressors, including the intercept, and we only choose p regressors ($p \leq k$) and obtain the residual sum of squares (RSS) using these p regressors, denoted as RSS_p. The C_p criterion is then defined as:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \tag{8.20}$$

where n is the number of observations, and $\hat{\sigma}^2$ is the estimator of the true variance. The expectation of C_p can be approximated as:

$$E(C_p) = \frac{(n-p)\sigma^2}{\sigma^2} - (n - 2p) \tag{8.21}$$

This means that in choosing a model according to the C_p criterion, one would look for a model that has a low C_p value, about equal to p. Following the principle of parsimony, a model with p regressors that gives a fairly good fit to the data would be chosen. In practice, one usually plots C_p against p, and an "adequate" model will show up as a point close to the $C_p = p$ line. For example, if two models are compared, Model A may be preferable to Model B if it is closer to the $C_p = p$ line, indicating a better balance between fit and complexity.

8.5. Self-Assessment Questions

- Can you explain the difference between in-sample and out-of-sample forecasting? Why is it important to consider both when selecting a model?
- Can you list and describe the six common types of specification errors? Can you provide an example of a situation where each might occur?
- What are the potential consequences of underfitting or overfitting a model? How can these be detected and corrected?
- Can you explain the steps involved in the Durbin-Watson “d” statistics test for specification error? When would you use this test?
- Can you describe the errors in the measurement of dependent variable "Y" and explanatory variable "X"? How can these errors impact the model?
- Can you define nested and non-nested models? Can you provide an example of each and explain how they differ?
- Can you explain the J test used for non-nested hypotheses? What are its steps, and what are some potential problems with this test?
- Can you compare and contrast the AIC, SIC, and Mallows's C_p criterion? When might you prefer one over the others?
- Given a dataset and a specific research question, how would you approach model selection, including the criteria and tests you would use?
- Imagine you are working with a model and discover a specification error. Can you outline the steps you would take to diagnose and correct the error?
- Can you explain the concept of adjusted R^2 and how it penalizes for adding more regressors? Why might it be preferred over R^2 in some cases?
- Can you provide an example from economics where non-nested models might be used to explain a phenomenon? How would you approach testing these models?
- Can you interpret a given plot of C_p against the number of regressors? How would you identify an “adequate” model from this plot?
- Given a real-world scenario, can you outline a comprehensive approach to model selection, testing, and validation, considering all the concepts covered in the Unit?

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Griffiths, W. E., Hill, R. C., & Judge, G. G. *Learning and Practicing Econometrics*, Wiley.
- Harvey, A. The Econometric Analysis of Time Series, Latest edition, The MIT Press, Cambridge.
- Hendry, D. F., & Richard, J. F. (1983). The Econometric Analysis of Economic Time Series, *International Statistical Review*, 51, pp. 3–33.
- Kmenta, J. *Elements of Econometrics*, Latest edition, Macmillan, New York, p. 431.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.

UNIT 09

SIMULTANEOUS EQUATION MODELS

Written By: **Dr. Muhammad Jamil**
Reviewed By: **Rizwan Ahmed Satti**

CONTENTS

Page Nos.

9.1. INTRODUCTION

In this Unit 09, the complex and multifaceted realm of simultaneous equation models is explored, delving into the very nature and structure that define these models. The Unit begins by elucidating the inherent characteristics of simultaneous equation models, laying the groundwork for a deeper understanding. It then navigates through the critical distinctions between endogenous and exogenous variables, providing insights into their roles and functions within the models. The exploration continues with an examination of structural equations and reduced-form equations, shedding light on their interplay and significance. A pivotal section on the identification problem introduces the reader to the challenges and intricacies involved in model identification, followed by a comprehensive discussion on various methods of identification, including the order and rank conditions. The Unit then transitions into the methodologies of estimation, offering a detailed analysis of techniques such as Recursive Models, Ordinary Least Squares, Indirect Least Square (ILS), and Two Stage Least Square (2SLS). Finally, the Unit concludes with a reflective look at the limitations of dynamic analysis, providing a sobering perspective on the challenges and constraints that practitioners may encounter. Overall, this Unit serves as a robust guide to the multifarious aspects of simultaneous equation models, blending theoretical insights with practical methodologies.

9.2. OBJECTIVES

After thorough study of the unit, you will be able to:

- **understand the Nature of Simultaneous Equation Models:** Grasp the fundamental characteristics and principles that define simultaneous equation models, laying a solid foundation for further exploration.
- **differentiate between Endogenous and Exogenous Variables:** Learn to identify and distinguish between these two types of variables, understanding their roles, functions, and implications within the models.
- **comprehend Structural and Reduced-Form Equations:** Develop an understanding of the interplay between structural equations and reduced-form equations and recognize their significance in the modeling process.
- **master the Identification Problem:** Gain insights into the challenges of model identification, including the complexities and intricacies involved in this critical aspect of modeling.
- **apply Methods of Identification:** Acquire the skills to apply various methods of identification, including the order and rank conditions, to ensure the validity and reliability of the models.

- **implement Methods of Estimation:** Learn and practice various estimation techniques such as Recursive Models, Ordinary Least Squares, Indirect Least Square (ILS), and Two Stage Least Square (2SLS), understanding their applications and limitations.
- **evaluate Limitations of Dynamic Analysis:** Reflect on the constraints and challenges of dynamic analysis within the context of simultaneous equation models, fostering a critical and nuanced perspective.
- **integrate Theory with Practice:** Synthesize theoretical insights with practical methodologies, applying the knowledge gained to real-world scenarios and problems.
- **cultivate Critical Thinking and Analytical Skills:** Encourage the development of critical thinking and analytical skills through the examination of complex concepts, methodologies, and challenges in the field of simultaneous equation models.
- **prepare for Advanced Study:** Equip students with the foundational knowledge and skills necessary for more advanced study in econometrics and related fields, fostering a lifelong learning attitude.

By achieving these objectives, students will be well-prepared to navigate the multifaceted world of simultaneous equation models, with a robust understanding of the underlying principles, methodologies, and practical applications.

9.3. Major Topics

- The Nature of the Simultaneous Equation Models
- Endogenous and Exogenous Variables
- Structural Equations and Reduced form Equations
- The Identification Problem
- Methods of Identification
- Methods of Estimations (OLS, ILS, 2SLS)
- Limitations of Dynamic Analysis

9.4. Summary of the Units

9.4.1. The Nature of the Simultaneous Equation Models

In the study of statistical models, simultaneous-equation models represent a significant departure from single-equation models. While single-equation models focus on a unidirectional cause-and-effect relationship between a dependent variable Y and one or more explanatory variables X, simultaneous-equation models recognize that this relationship may be bidirectional.

The nature of simultaneous-equation models is such that a set of variables can be determined simultaneously by the remaining set of variables. This leads to a system of equations where each equation represents one of the mutually or jointly dependent (endogenous) variables. Here's a hypothetical example:

$$Y_{1i} = \beta_{10} + \beta_{12}Y_{2i} + \gamma_{11}X_{1i} + u_{1i} \quad (9.1)$$

$$Y_{2i} = \beta_{20} + \beta_{21}Y_{1i} + \gamma_{21}X_{1i} + u_{2i} \quad (9.2)$$

In this system, Y_1 and Y_2 are mutually dependent endogenous variables, and X_1 is an exogenous variable. The variables u_1 and u_2 are stochastic disturbance terms.

A critical aspect of simultaneous-equation models is that estimating the parameters of a single equation without considering the others in the system can lead to problems. Specifically, if the method of ordinary least squares (OLS) is applied to each equation individually, disregarding the simultaneous nature of the system, the estimators may be biased and inconsistent. This inconsistency arises if the stochastic explanatory variables are not distributed independently of the stochastic disturbance terms.

In the given example, unless it can be shown that Y_2 in Eq. (9.1) is distributed independently of u_1, and Y_1 in Eq. (9.2) is distributed independently of u_2, the application of classical OLS will lead to inconsistent estimates.

In summary, simultaneous-equation models recognize the complex, bidirectional relationships that can exist between variables. They require special methods of estimation that take into account the entire system of equations, as the direct application of standard techniques like OLS can lead to biased and inconsistent results. The Unit further explores examples of these models and introduces specialized methods developed to handle them, recognizing the inherent challenges in estimating such systems.

9.4.2. Endogenous and Exogenous Variables

In simultaneous-equation models, the distinction between dependent and explanatory variables becomes more complex, leading to the concepts of endogenous and exogenous variables. Here's a summary of these concepts:

- **Endogenous Variables:** These are variables that are determined within the system of equations. They are mutually dependent and are affected by other variables in the system. In the example given in the text:

  $$Y_{1i} = \beta_{10} + \beta_{12}Y_{2i} + \gamma_{11}X_{1i} + u_{1i} \quad (9.1)$$
  $$Y_{2i} = \beta_{20} + \beta_{21}Y_{1i} + \gamma_{21}X_{1i} + u_{2i} \quad (9.2)$$

  In this system, Y_1 and Y_2 are endogenous variables, as they are determined by each other and the exogenous variable X_1. If these variables are treated as stochastic and not distributed independently of the stochastic disturbance terms, the application of ordinary least squares (OLS) to these equations individually will lead to inconsistent estimates.

- **Exogenous Variables:** These are variables that are determined outside the system of equations. They are not affected by the endogenous variables within the system. In the example above, X_1 is an exogenous variable. It is not influenced by the endogenous variables Y_1 and Y_2, and its value is predetermined.

- **Lagged Variables:** The text also mentions the concept of lagged variables, which can be either exogenous or endogenous. For example, $X_{1(t-1)}$ is a lagged exogenous variable with a lag of one time period, and Y_{t-1} is a lagged endogenous variable with a lag of one time period. Lagged variables are considered predetermined, as their values are not determined by the model in the current time period.

In summary, the concepts of endogenous and exogenous variables are central to the understanding of simultaneous-equation models. Endogenous variables are determined within the system and may influence each other, while exogenous variables are determined outside the system and are not influenced by the endogenous variables. Proper classification and understanding of these concepts

are crucial for the correct estimation and interpretation of simultaneous-equation models.

9.4.3. Structural Equations and Reduced Form Equations

In the study of economic relationships, it is often essential to consider the interconnected nature of variables, where changes in one variable can influence and be influenced by changes in others. This complexity is captured through simultaneous-equation models, a powerful tool that allows for the modeling of mutual dependencies between variables.

- **Structural Equations:** Structural equations represent the theoretical relationships between variables as derived from economic theory or other substantive considerations. They are the fundamental equations that describe how the endogenous variables are determined within the system. Structural equations typically include both endogenous and exogenous variables, as well as stochastic error terms. Equations (9.1) and (9.2) are structural equations, as they represent the theoretical relationships between the endogenous variables Y_1 and Y_2, the exogenous variable X_1, and the error terms u_1 and u_2.
- **Reduced Form Equations:** Reduced-form equations are derived from the structural equations by solving them for the endogenous variables in terms of only the exogenous variables and the error terms. They do not contain any endogenous variables on the right-hand side. Reduced-form equations provide the statistical relationships that can be estimated directly from the data, without any assumptions about the structural parameters.

  If we were to solve the structural equations above for Y_1 and Y_2 in terms of X_1, we would obtain the reduced-form equations. These equations would express Y_1 and Y_2 solely as functions of the exogenous variable X_1 and new error terms, which would be combinations of the original error terms u_1 and u_2.

Structural equations represent the underlying theoretical relationships in a simultaneous-equation model, while reduced-form equations provide the empirical relationships that can be estimated directly. Understanding both forms is essential for the proper estimation and interpretation of simultaneous-equation models.

9.4.4. The Identification Problem

The identification problem in simultaneous-equation models is a critical aspect of econometric analysis. It refers to the challenge of determining whether numerical estimates of the parameters of a structural equation can be obtained from the

estimated reduced-form coefficients. This problem can manifest in three ways: under identification, exact identification, and over identification.

- **Under identification:** An equation is considered under identified or unidentified if it is impossible to obtain estimates of the structural parameters. No matter how extensive the data, the structural parameters cannot be estimated. Most simultaneous-equation systems in economics and finance are over identified rather than under identified, so under identification is often not a major concern.

- **Exact Identification:** An equation is said to be exactly identified if unique numerical values of the structural parameters can be obtained. This means that there is a one-to-one correspondence between the structural and reduced-form parameters, allowing for precise estimation.

- **Over identification:** An equation is considered over identified if more than one numerical value can be obtained for some of the parameters of the structural equations. In this case, there may be several estimates of one or more structural coefficients, leading to ambiguity in the interpretation of the model.

The identification problem arises because different sets of structural coefficients may be compatible with the same set of data. This can make it difficult to determine which particular hypothesis or model is being investigated. The circumstances under which each of these cases occurs can be complex, and special methods have been developed to handle them.

The identification problem is fundamental in simultaneous-equation models, as it precedes the problem of estimation. If an equation is identified, it can be either just identified or over identified. In the former case, unique values of structural coefficients can be obtained; in the latter, there may be more than one value for one or more structural parameters.

9.4.5. Methods of Identification

The section on Methods of Identification in simultaneous equation models from the book delves into the systematic routine of determining the identification of an equation in a system of simultaneous equations. This process can be time-consuming and laborious, but the introduction of order and rank conditions lightens the task.

9.4.5.1. The Order Condition of Identifiability

The order condition is a necessary but not sufficient condition for identification. It is defined as:

In a model of M simultaneous equations, in order for an equation to be identified, the number of predetermined variables excluded from the equation must not be less than the number of endogenous variables included in that equation less 1, that is

$$K - k \geq m - 1 \qquad (9.3)$$

Where, K is number of predetermined variables in the model, including the intercept, k is number of predetermined variables in each equation, m is number of endogenous variables in a given equation and M is number of endogenous variables in the model.

In the above condition if $K - k = m - 1$, the equation is just identified, but if $K - k > m - 1$, it is over identified. This condition is essential for understanding whether an equation is identified, but it may not be enough on its own.

9.4.5.2. The Rank Condition of Identifiability

The rank condition of identification is a necessary and sufficient condition that goes beyond the order condition in simultaneous equation models. While the order condition is necessary, it may not be sufficient for identification. The rank condition ensures that the predetermined variables excluded from a particular equation but present in the model are all independent, allowing a one-to-one correspondence between the structural coefficients (the β's) and the reduced-form coefficients. This ensures that the structural parameters can be estimated from the reduced-form coefficients.

In essence, the rank condition of identification ensures that the equation is identified only if the variables excluded from the equation influence the other equations in the system. It provides a more robust and definitive criterion for identification, ensuring that the model's parameters can be estimated with confidence.

In a model containing M equations in M endogenous variables, an equation is identified if and only if at least one nonzero determinant of order $(M - 1)(M - 1)$ can be constructed from the coefficients of the variables (both endogenous and predetermined) excluded from that particular equation but included in the other equations of the model.

The rank condition is a method used to determine the identifiability of a structural equation in a system of simultaneous equations. Here's how to apply the rank condition:

- **Step 1 – Tabulate the System:** Write down the system of equations in a tabular form.

- **Step 2 – Strike out Row Coefficients:** In the row where the equation under consideration appears, strike out the coefficients.
- **Step 3 – Strike out Corresponding Columns:** Also, strike out the columns corresponding to those coefficients in step (2) that are nonzero.
- **Step 4 – Form Matrices:** The remaining entries in the table will give only the coefficients of the variables included in the system but not in the equation under consideration. From these entries, form all possible matrices of order $M-1$ and obtain the corresponding determinants.
- **Step 5 – Determine Identifiability:** If at least one nonvanishing or nonzero determinant can be found, the equation in question is identified (either just or over-identified). The rank of the matrix in this case is exactly equal to $M-1$. If all the possible $(M-1)(M-1)$ determinants are zero, the rank of the matrix is less than $M-1$, and the equation is not identified.

The general principles of identifiability derived from the rank condition are as follows:

- Over-Identified: If $K-k > m-1$ and the rank of the A matrix is $M-1$, the equation is over-identified.
- Exactly Identified: If $K-k > m-1$ and the rank of the matrix A is $M-1$, the equation is exactly identified.
- Under-Identified: If $K-k \geq m-1$ and the rank of the matrix A is less than $M-1$, the equation is under-identified.
- Unidentified: If $K-k < m-1$, the structural equation is unidentified. The rank of the A matrix in this case is bound to be less than $M-1$.

This method ensures that the structural parameters of the model can be accurately estimated, considering the simultaneous relationships between variables.

9.4.6. Methods of Estimations

In the context of simultaneous-equation models with M endogenous variables, there are two main approaches to estimate the structural equations: single-equation methods (limited information methods) and system methods (full information methods). Single-equation methods estimate each equation individually, considering only the restrictions placed on that specific equation. System methods, on the other hand, estimate all equations simultaneously, considering all restrictions across the system. Ideally, the systems method, such as the full information maximum likelihood (FIML) method, should be used to preserve the spirit of

simultaneous-equation models. However, in practice, these methods are not commonly used due to enormous computational burdens, highly nonlinear solutions in the parameters, and sensitivity to specification errors. Even with high-speed computers, the computations for large models can be a daunting task, and any error in one equation can affect the entire system.

In practice, therefore, single-equation methods are often used. Following are the methods of estimation based on single equation methods.

9.4.6.1. Recursive Models and Ordinary Least Squares

In the context of simultaneous equations, the Ordinary Least Squares (OLS) method is generally inappropriate due to the interdependence between the stochastic disturbance term and the endogenous explanatory variables. If applied incorrectly, the estimators are biased and inconsistent. However, there is an exception where OLS can be appropriately applied: the case of recursive, triangular, or causal models.

Consider a three-equation system:

$$Y_{1t} = \beta_{10} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + u_{1t} \quad (9.4)$$

$$Y_{2t} = \beta_{20} + \beta_{21}Y_{1t} + \gamma_{21}X_{1t} + \gamma_{22}X_{2t} + u_{2t} \quad (9.5)$$

$$Y_{3t} = \beta_{30} + \beta_{31}Y_{1t} + \beta_{32}Y_{2t} + \gamma_{31}X_{1t} + \gamma_{32}X_{2t} + u_{3t} \quad (9.6)$$

Here, the disturbances are such that:

$$cov(u_{1t}, u_{2t}) = cov(u_{1t}, u_{3t}) = cov(u_{2t}, u_{3t}) = 0$$

The first equation contains only exogenous variables uncorrelated with the disturbance term u_{1t}, so OLS can be applied. In the second equation, OLS can also be applied, provided Y_{1t} and u_{2t} are uncorrelated, which is true since u_1, affecting Y_1, is uncorrelated with u_2. The same logic extends to the third equation, allowing OLS to be applied to each equation separately.

In the recursive system, there is no interdependence among the endogenous variables; each equation exhibits unilateral causal dependence. For example, Y_1 affects Y_2, but Y_2 doesn't affect Y_1. This lack of mutual influence allows for the application of OLS, making it clear that there is no simultaneous-equation problem in this situation. The structure of these systems leads to the name "causal models."

9.4.6.2. Indirect Least Square (ILS)

The method of Indirect Least Squares (ILS) is used to obtain estimates of the structural coefficients from the Ordinary Least Squares (OLS) estimates of the reduced-form coefficients for a just or exactly identified structural equation. The ILS method involves the following three steps:

- **Step 1 – Obtain the Reduced-Form Equations:** The first step is to obtain the reduced-form equations from the structural equations. In these reduced-form equations, the dependent variable in each equation is the only endogenous variable and is a function solely of the predetermined variables (exogenous or lagged endogenous) and the stochastic error term(s).
- **Step 2 – Apply OLS to the Reduced-Form Equations:** In the second step, OLS is applied to the reduced-form equations individually. This operation is permissible since the explanatory variables in these equations are predetermined and hence uncorrelated with the stochastic disturbances. The estimates obtained through this step are consistent.
- **Step 3 – Obtain Estimates of the Original Structural Coefficients:** The final step involves obtaining estimates of the original structural coefficients from the estimated reduced-form coefficients obtained in Step 2. If an equation is exactly identified, there is a one-to-one correspondence between the structural and reduced-form coefficients, allowing unique estimates of the former to be derived from the latter.

The name "Indirect Least Squares" reflects the fact that the structural coefficients, which are often the primary object of inquiry, are obtained indirectly from the OLS estimates of the reduced-form coefficients. This method provides a systematic way to translate the information contained in the reduced-form equations into insights about the underlying structural relationships.

The Indirect Least Squares (ILS) estimators possess certain properties that are inherited from the reduced-form estimators. These include consistency and asymptotic efficiency, meaning that as the sample size increases indefinitely, the estimators converge to their true values and achieve the lowest possible variance. However, in small samples, the estimators may not necessarily be unbiased. This means that the expected value of the estimators may not equal the true population parameters. For instance, in the case of the supply function, the ILS estimators are biased in small samples, but this bias disappears as the sample size increases, demonstrating the property of consistency. Therefore, while the ILS method has desirable asymptotic properties, its performance in small samples may be less satisfactory due to potential bias.

9.4.6.3. Two Stage Least Square (2SLS)

The Two-Stage Least Squares (2SLS) method is a sophisticated technique employed to address the challenges of estimating simultaneous equation models. It's particularly useful when dealing with endogeneity issues, where ordinary least squares (OLS) would yield biased and inconsistent estimators.

To illustrate 2SLS, consider the following system of equations:

$$Y_{1t} = \beta_{10} + \beta_{12}Y_{2t} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + u_{1t} \quad (9.7)$$
$$Y_{2t} = \beta_{20} + \beta_{21}Y_{1t} + \gamma_{23}X_{3t} + \gamma_{24}X_{4t} + u_{2t} \quad (9.10)$$

The 2SLS method is implemented through a sequence of well-defined steps, as follows:

- **First Stage:** In the initial stage, each endogenous explanatory variable in the structural equation under consideration is regressed on all the exogenous variables in the system, including those that do not appear in the equation being estimated. This leads to the estimation of the predicted values of the endogenous explanatory variables. Mathematically, this can be represented as:

  $$Y_{1t} = \pi_{10} + \pi_{11}X_{1t} + \pi_{12}X_{2t} + \pi_{13}X_{3t} + \pi_{14}X_{4t} + u_{1t} \quad (9.11)$$
  $$Y_{2t} = \pi_{20} + \pi_{21}X_{1t} + \pi_{22}X_{2t} + \pi_{23}X_{3t} + \pi_{24}X_{4t} + u_{2t} \quad (9.12)$$

  Where Y_{1t} and Y_{2t} are the endogenous variable, X_{1t}, X_{2t}, X_{3t}, and X_{4t} are the exogenous variables.

- **Second Stage:** In the subsequent stage, the structural equation is estimated by replacing the endogenous explanatory variables with the predicted values obtained from the first stage. This is akin to applying OLS to the modified equation:

  $$Y_{1t} = \beta_{10} + \beta_{12}\hat{Y}_{2t} + \gamma_{11}X_{1t} + \gamma_{12}X_{2t} + u_{1t} \quad (9.13)$$
  $$Y_{2t} = \beta_{20} + \beta_{21}\hat{Y}_{1t} + \gamma_{23}X_{3t} + \gamma_{24}X_{4t} + u_{2t} \quad (9.14)$$

  Where $\hat{Y}_{1t}$ and $\hat{Y}_{2t}$ are the predicted value of the endogenous explanatory variable from the first stage.

The 2SLS method has been lauded for its ability to provide consistent estimators even when the endogenous explanatory variables are correlated with the error terms. However, it's worth noting that the estimators may not be efficient if the errors are heteroskedastic or correlated across equations. Several studies have expounded upon the properties and applications of 2SLS. For instance, the seminal work by Theil (1953) and Basmann (1957) has been instrumental in elucidating the theoretical underpinnings of this method.

In summary, the Two-Stage Least Squares method is a recondite yet powerful tool in econometric analysis, providing a robust solution to the challenges posed by simultaneous equation models. Its step-by-step approach allows for the consistent estimation of structural parameters, making it a preferred choice among researchers and practitioners alike.

9.4.7. Limitations of Dynamic Analysis

Here's a synthesis of the insights:

- **Data Quality and Model Formulation:** The results of research are only as good as the quality of the data. If the data quality is poor, the results may be unsatisfactory. The inability to formulate the model precisely due to weak underlying theory or lack of appropriate data can also be a limitation.
- **Computational Burden and Sensitivity to Errors:** Systems methods like full information maximum likelihood (FIML) are computationally intensive and can lead to highly nonlinear solutions. They are also very sensitive to specification errors, making single-equation methods often more practical.
- **Panel Data Challenges:** Despite the advantages of panel data, they pose several estimation and inference problems. Issues like heteroscedasticity, autocorrelation, and cross-correlation in individual units need to be addressed.
- **Limitations in Time Series Analysis:** In economic time series data, successive values tend to be highly correlated, leading to multicollinearity. This results in imprecise estimation and potential erroneous conclusions about statistical significance. The sequential search for the lag length also opens the researcher to the charge of data mining.
- **Challenges with VAR Models:** Vector autoregression (VAR) models, while emphasizing forecasting, are less suited for policy analysis. Choosing the appropriate lag length is a significant challenge, and estimating many parameters can consume a lot of degrees of freedom. Ensuring stationarity in all variables is also a strict requirement.
- **No Single Solution:** There may be more than one solution to a particular problem, and its often unclear which method is best. Multiple violations of the classical linear regression model may coexist, and there is no single test that will solve all problems simultaneously.
- **Limitations in Estimating Time-Invariant Variables:** In some situations, methods like the Least Squares Dummy Variable (LSDV) approach may not be able to identify the impact of time-invariant variables, making precise estimation difficult.

These limitations underscore the complexity and challenges in dynamic analysis, requiring careful consideration of methodological choices, data quality, and the underlying assumptions of the models. The recondite nature of these challenges emphasizes the need for robust methodological rigor and a nuanced understanding of the underlying econometric principles.

9.5. Self-Assessment Questions

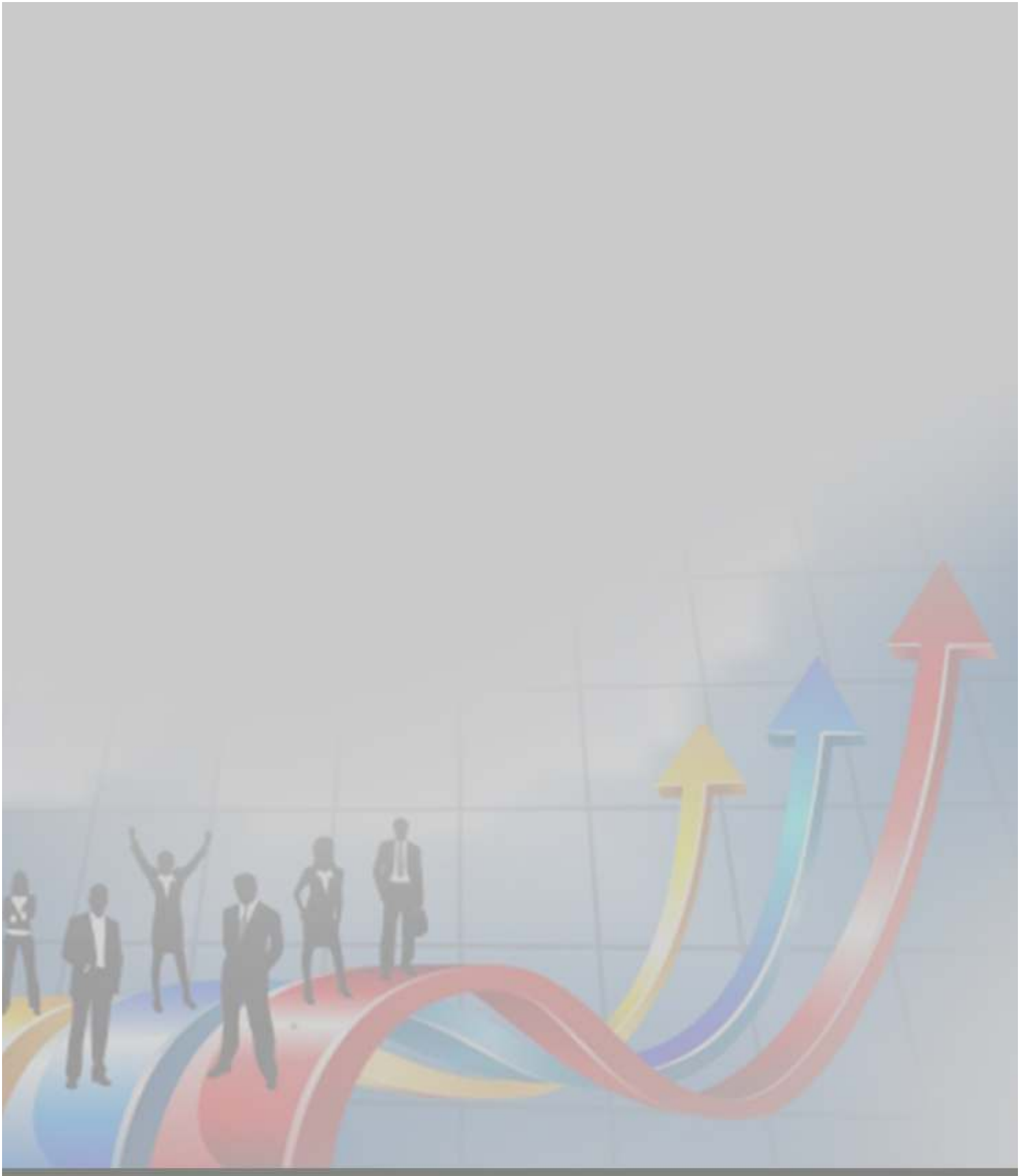
- What are simultaneous equation models, and why are they important in econometric analysis? Provide an example to illustrate your understanding.
- Define endogenous and exogenous variables. How do they differ, and what roles do they play in simultaneous equation models?
- Explain the relationship between structural equations and reduced-form equations. Why are both forms significant in modeling?
- What is the identification problem in the context of simultaneous equation models? Why is it considered a challenge?
- Describe the order condition of identifiability. How does it contribute to the identification of a model?
- Explain the rank condition of identifiability and outline the steps involved in applying it.
- How can recursive models be estimated using Ordinary Least Squares (OLS)? Provide an example to illustrate the process.
- Summarize the Indirect Least Square (ILS) method and its steps. When is it appropriate to use this method?
- Describe the Two Stage Least Square (2SLS) method. What are its key features, and how is it implemented?
- Discuss the limitations of dynamic analysis in simultaneous equation models. Provide examples or scenarios where these limitations might be evident.
- How do the concepts of endogenous and exogenous variables, structural and reduced-form equations, and various estimation methods interrelate in the context of simultaneous equation models?
- Reflect on a real-world application of simultaneous equation models that you have studied. What methods were used, and how were the challenges and limitations addressed?
- If you were to conduct a study using simultaneous equation models, what approach would you take? Outline your methodology, including the identification and estimation methods you would employ.

Textbooks & Supplies

- Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill Education
- Maddala, G. S. Econometrics, McGraw-Hill Company.
- Koutsoyiannis, A. Theory of Econometrics. Latest Edition, McMillan.

Additional Readings

- Basmann, R. L. (1957). A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation. *Econometrica*, 25(1), 77-83.
- Griffiths, W. E., Hill, R. C., & Judge, G. G. *Learning and Practicing Econometrics*, Wiley.
- Kmenta, J. *Elements of Econometrics*, Latest edition, Macmillan, New York.
- Theil, H. (1953). Repeated Least-Squares Applied to Complete Equation Systems, The Hague: The Central Planning Bureau, The Netherlands.
- Wooldridge, J. M. Introductory Econometrics: A Modern Approach. South-Western Cengage Learning.



Fundamental of Econometrics