BACHELOR OF BUSINESS ADMINISTRATION
(FINANCIAL INVESTMENT ANALYSIS)

BASICS OF ECONOMETRICS

DISCIPLINE SPECIFIC CORE COURSE (DSC-10)
SEMESTER - IV COURSE CREDIT -4

(FOR LIMITED CIRCULATION ONLY)

DEPARTMENT OF DISTANCE AND CONTINUING EDUCATION
UNIVERSITY OF DELHI

AS PER THE UGCF-2022 AND NATIONAL EDUCATION POLICY 2020

1st Edition: 2024

E-mail: ddceprinting@col.du.ac.in
financialstudies@col.du.ac.in

Published by:
Department of Distance and Continuing Education
Campus of Open Learning/School of Open Learning,
University of Delhi, Delhi-110007

Printed by:
School of Open Learning, University of Delhi

Internal Review Committee

Mr. Mukesh Kumar

Corrections/Modifications/Suggestions proposed by Statutory Body, DU/ Stakeholder/s in the Self Learning Material (SLM) will be incorporated in the next edition. However, these corrections/modifications/suggestions will be uploaded on the website https://sol.du.ac.in.

Any feedback or suggestions can be sent to the email-feedback.slm@col.du.ac.in.

Printed at: **Taxmann Publications Pvt. Ltd., 21/35, West Punjabi Bagh New Delhi - 110026 (500 Copies, 2024)**

Department of Distance & Continuing Education, Campus of Open Learning, School of Open Learning, University of Delhi

Contents

PAGE

UNIT-I

Lesson 1: Introduction to Econometrics

1.1 Learning Objectives 3
1.2 Introduction to Econometrics 4
1.3 Rationale Behind Having a Separate Discipline 4
1.4 Methodology of Econometrics 5
1.5 Introduction to Econometric Softwares 12
1.6 Summary 17
1.7 Answers to In-Text Questions 17
1.8 Self-Assessment Questions 18
1.9 References 18
1.10 Suggested Readings 18

Lesson 2: Regression Models: Assumption, Properties, Estimation and Hypothesis Testing

2.1 Learning Objectives 19
2.2 Introduction 20
2.3 Model Specification 21
2.4 Estimation 24
2.5 Testing 25
2.6 Multicollinearity and Variable Inflation Factor (VIF) 30
2.7 Residual Analysis 34
2.8 Answers to In-Text Questions 40
2.9 Self-Assessment Questions 41

PAGE

2.10 Summary 42
2.11 References 43
2.12 Suggested Readings 43

UNIT-II

Lesson 3: Violations of Classical Assumptions 1

3.1 Learning Objectives 47
3.2 Introduction 48
3.3 Multicollinearity 48
3.4 Heteroscedasticity 57
3.5 Summary 66
3.6 Answers to In-Text Questions 67
3.7 Self-Assessment Questions 67
3.8 References 68
3.9 Suggested Readings 68

Lesson 4: Violations of Classical Assumptions 2

4.1 Learning Objectives 69
4.2 Introduction 70
4.3 Autocorrelation 70
4.4 Specification Errors 83
4.5 Summary 89
4.6 Answers to In-Text Questions 89
4.7 Self-Assessment Questions 90
4.8 References 90
4.9 Suggested Readings 90

UNIT-III

Lesson 5: Goodness of Fit

5.1 Learning Objectives 93
5.2 Introduction 94

PAGE

5.3 What is Goodness of Fit? 94
5.4 Test/Statistics Used for Goodness of Fit 94
5.5 R Square/R^2 99
5.6 Adjusted R Square/*Adjusted* R^2 101
5.7 Standard Error of the Model 102
5.8 Conceptual Understanding of AIC, BIC and SIC 104
5.9 Calculation and Comparison of AIC, BIC and SIC 106
5.10 Summary 109
5.11 Answers to In-Text Questions 110
5.12 Self-Assessment Questions 110
5.13 References 110
5.14 Suggested Readings 111

UNIT-IV

Lesson 6: Dummy Variables and Panel Data Regression Models

6.1 Learning Objectives 116
6.2 Introduction 116
6.3 Concept of Dummy Variables 117
6.4 Types of Dummy Variables 118
6.5 Use of Dummy Variables to Model Qualitative/Binary/Structural Changes 123
6.6 Other Functional Forms of Dummy Variables 124
6.7 Response Regression Models 125
6.8 Panel Data Regression Model 128
6.9 Different Methods of Panel Data Estimation 132
6.10 Summary 135
6.11 Answers to In-Text Questions 136
6.12 Self-Assessment Questions 136
6.13 References 136
6.14 Suggested Readings 137
Glossary 139

UNIT - I

LESSON

1

Introduction to Econometrics

Dr. Ruhee Mittal
Department of Economics
School of Open Learning
University of Delhi
Email-Id: ruhee.mittal@sol-du.ac.in

STRUCTURE

1.1 *Learning Objectives*
1.2 *Introduction to Econometrics*
1.3 *Rationale Behind Having a Separate Discipline*
1.4 *Methodology of Econometrics*
1.5 *Introduction to Econometric Softwares*
1.6 *Summary*
1.7 *Answers to In-Text Questions*
1.8 *Self-Assessment Questions*
1.9 *References*
1.10 *Suggested Readings*

1.1 Learning Objectives

- Comprehend the significance of econometrics in bridging economic theory with empirical analysis through statistical inference.
- Learn to assess and interpret the statistical models utilized in econometrics to analyze economic phenomena.
- Acquire the skills to apply econometric tools for forecasting and policy analysis in real-world economic scenarios.

Notes

1.2 Introduction to Econometrics

Econometrics, originally rooted in the notion of "economic measurement," extends beyond mere quantification. It embodies a fusion of mathematical statistics, economic theory, and empirical data analysis. This interdisciplinary field applies mathematical and statistical tools to validate economic models, interpret real-world phenomena, and derive quantitative insights. At its core, econometrics strives to empirically uncover economic laws by amalgamating theoretical frameworks with observed data. The craft of an econometrician lies in formulating assumptions that are both specific and realistic, leveraging available data optimally. In essence, econometrics acts as a bridge, connecting economic theory and real-world measurements through the prism of statistical inference, aiming to dispel misconceptions about the emptiness of economic theories and their various interpretations.

1.3 Rationale Behind Having a Separate Discipline

Econometrics mixes economics, math, and stats. But why focus on it separately? Well, economic theory tells us how things might work in the economy – like when prices drop, people usually buy more. But it doesn't give us specific numbers. That's where econometrics steps in, using math and stats to find those exact numbers. While math economics turns theories into equations, econometrics tests these equations using real data. Economic stats collect and organize this data, which econometricians then use to test economic theories. However, dealing with economic data isn't like controlled experiments in a lab; it needs special skills and methods because this data isn't controlled. Econometricians work with observed data, like meteorologists and they desire to understand the nature and structure of the data.

For example, economic theory suggests that when the price of a product decreases, the demand for that product increases. This theory is qualitative—it tells us about the relationship between price and demand but doesn't specify exactly how much the demand will change when the price changes.

Econometrics steps in to quantify this relationship using real-world data. For instance, let's say the theory predicts that a 10% decrease in the price of a certain product will lead to a 20% increase in its demand. Econometricians collect and analyze data on the prices and quantities

Notes

sold for this product across different periods and locations. By applying mathematical and statistical tools to this data, they aim to confirm or refute this predicted relationship between price and demand.

Using statistical methods, they might find that indeed, on average, a 10% drop in price corresponds to around a 20% increase in demand, validating the theory quantitatively. This process of analyzing real data to test and quantify the predictions made by economic theory is the essence of what econometrics does.

1.4 Methodology of Econometrics

The methodology of econometrics encompasses various approaches, with the traditional or classical method being the primary approach used in empirical research across economics and other social sciences. This classical approach follows these steps:

- Formulating a theory or hypothesis.
- Defining the mathematical representation of the theory.
- Developing the statistical or econometric model.
- Collecting relevant data.
- Estimating the parameters within the econometric model.
- Testing the formulated hypotheses.
- Making forecasts or predictions.
- Applying the model for control or policy purposes.

To illustrate these steps, let's examine the application of these principles using the widely recognized Keynesian theory of consumption.

1. **Formulating a Theory or Hypothesis:** Let's begin by articulating a theory or hypothesis about an economic phenomenon.

   For instance, according to Keynes the theory of consumption that individuals tend to boost their spending as their income rises, but not to the same extent as the income increases itself. He essentially posited that the Marginal Propensity to Consume (MPC), representing the rate at which spending changes for every unit increase in income (e.g., a dollar), stands above zero yet falls short of reaching 1.

Notes

2. **Defining the Mathematical Representation of the Theory:** While Keynes suggested a positive link between consumption and income, he didn't specify the exact nature of their functional relationship. To simplify, a mathematical economist might propose the following form of the Keynesian consumption function:

$$Y = \beta_1 + \beta_2 X \text{ where } 0 < \beta_2 < 1 \qquad (1.1)$$

In this equation, Y represents consumption expenditure, X stands for income, and the parameters β_1 and β_2 act as the intercept and slope coefficients, respectively. The slope coefficient β_2 serves as a measure of the MPC. Equation 1.1 can be depicted as shown in Figure 1.1, illustrating a linear relationship between consumption and income. This equation, known as the consumption function in economics, presents a mathematical model demonstrating the connection between consumption and income.

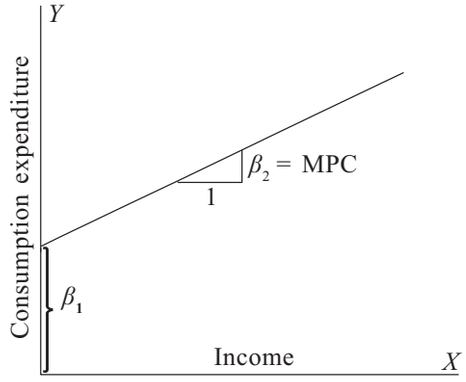


Figure 1.1

(Source: Gujarati, 2004)

In Eq. (1.1), the variable on the left side signifies the dependent variable, while the variable(s) on the right side denotes the independent or explanatory variable(s). Hence, in the Keynesian consumption function, Eq. (1.1), consumption (expenditure) serves as the dependent variable, whereas income acts as the explanatory variable.

3. **Developing the Statistical or Econometric Model:** The consumption function represented purely by the mathematical equation in Eq. (1.1) holds limited interest for the econometrician. This model assumes an exact or deterministic link between consumption and income. However, real-world economic relationships tend to be imprecise.

Notes

For instance, if data were collected on consumption expenditure and disposable income from, let's say, 500 American families and plotted on a graph, they wouldn't perfectly align with the straight line described by Eq. (I.3.1). This divergence occurs because factors beyond income, such as family size, ages of family members, religious beliefs, and more, also influence consumption.

To account for these imprecise relationships between economic variables, the econometrician adjusts the deterministic consumption function in Eq. (1.1) by introducing the following modification:

$$Y = \beta_1 + \beta_2 X + u \qquad (1.2)$$

Here, u is termed as the disturbance or error term, represents a random (stochastic) variable possessing well-defined probabilistic properties. This error term u encapsulates various unaccounted factors impacting consumption.

Equation (1.2) serves as an example of an econometric model, specifically, a linear regression model, which constitutes the primary focus of this book. This econometric model for consumption posits that while the dependent variable Y (consumption) has a linear relationship with the explanatory variable X (income), this relationship isn't exact; it varies across individuals.

A visual representation of the econometric consumption function is depicted in Figure 1.2.

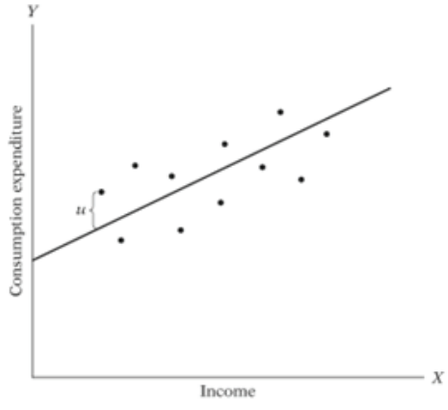


Figure 1.2

(Source: Gujarati, 2004)

Notes

4. **Collecting Relevant Data:** To derive the numerical values of β_1 and β_2 and in the econometric model illustrated by Eq. (1.2), data becomes imperative. While we'll delve deeper into the significance of data for economic analysis in the following chapter, let's presently focus on the data provided in Table 1.1, pertaining to the U.S. economy spanning the years 1960–2005. The variable Y in this table represents the total Personal Consumption Expenditure (PCE) for the entire economy, while X denotes the gross domestic product (GDP), a measure of overall income, both quantified in billions of 2000 dollars. Thus, these figures are articulated in "real" terms, denoting constant (2000) prices. The graphical representation of this data is depicted in Figure 1.3 (compare with Figure 1.2). For the moment, disregard the line portrayed within the figure.

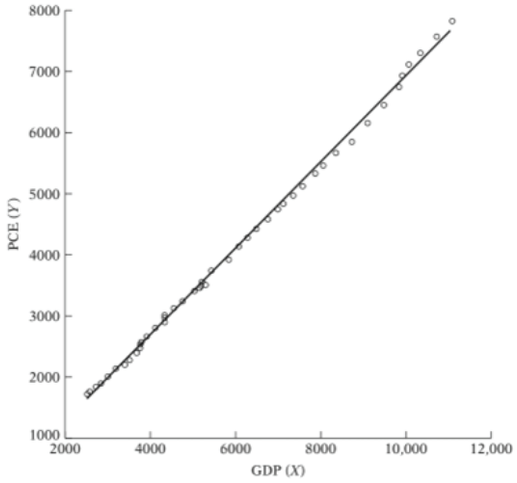


Figure 1.3: Figure Showing Personal Consumption Expenditure (Y) in Relation to GDP (X), 1960-2005, in Billions of 2000 Dollars

(Source: Gujarati, 2004)

Figure 1.3 depicts that the regression line is closely aligned with the data points, indicating a strong fit. Within the period from 1960 to 2005, the slope coefficient (representing the MPC) was approximately 0.72. This suggests that, on average during this time-frame, a one-dollar rise in real income correlated with an increase of around 72 cents in real consumption expenditure. It's essential to note the "on average" aspect since the connection between consumption and income isn't precise – as visible in Figure 1.3, not all data points

Notes

precisely follow the regression line. In simpler terms, based on our data, the average or mean consumption expenditure increased by about 72 cents for each one-dollar rise in real income.

5. **Estimating the Parameters within the Econometric Model:** With the data in hand, our subsequent objective is to calculate the parameters within the consumption function. These numerical values attributed to the parameters offer empirical substance to the consumption function. The detailed process of parameter estimation will be covered in Chapter 3. However, it's worth noting that the primary method employed for obtaining these estimates is the statistical procedure known as regression analysis. Employing this technique and utilizing the data provided in Table 1.1, we derive the subsequent estimates for β_1 and β_2: −299.5913 and 0.7218, respectively. Consequently, the resultant estimated consumption function reads as follows:

$$\hat{Y}_t = -299.5913 + 0.7218\ X_t \qquad (1.3)$$

Table 1.1

Year	PCE(Y)	GDP(X)
1960	1597.4	2501.8
1961	1630.3	2560.0
1962	1711.1	2715.2
1963	1781.6	2834.0
1964	1888.4	2998.6
1965	2007.7	3191.1
1966	2121.8	3399.1
1967	2185.0	3484.6
1968	2310.5	3652.7
1969	2396.4	3765.4
1970	2451.9	3771.9
1971	2545.5	3898.6
1972	2701.3	4105.0
1973	2833.8	4341.5
1974	2812.3	4319.6
1975	2876.9	4311.2
1976	3035.5	4540.9
1977	3164.1	4750.5
1978	3303.1	5015.0
1979	3383.4	5173.4
1980	3374.1	5161.7
1981	3422.2	5291.7
1982	3470.3	5189.3
1983	3668.6	5423.8
1984	3863.3	5813.6
1985	4064.0	6053.7
1986	4228.9	6263.6
1987	4369.8	6475.1
1988	4546.9	6742.7
1989	4675.0	6981.4
1990	4770.3	7112.5
1991	4778.4	7100.5
1992	4934.8	7336.6
1993	5099.8	7532.7
1994	5290.7	7835.5
1995	5433.5	8031.7
1996	5619.4	8328.9
1997	5831.8	8703.5
1998	6125.8	9066.9
1999	6438.6	9470.3
2000	6739.4	9817.0
2001	6910.4	9890.7
2002	7099.3	10048.8
2003	7295.3	10301.0
2004	7577.1	10703.5
2005	7841.2	11048.6

(Source: Gujarati, 2004)

Notes

6. **Testing the Formulated Hypotheses:** Supposing that the fitted model provides a reasonably accurate representation of reality, it becomes crucial to establish appropriate criteria for assessing whether the estimates, as presented in Eq. (1.3), align with the expectations of the theory being tested. According to "positive" economists like Milton Friedman, **a theory or hypothesis** that lacks verifiability through empirical evidence might not be considered a viable component of scientific inquiry.

   As previously mentioned, Keynes anticipated a positive but less than 1 MPC. In our scenario, the MPC was calculated to be roughly 0.72. However, before embracing this discovery as confirmation of Keynesian consumption theory, it's imperative to investigate whether this estimation significantly falls below unity, ensuring that it isn't merely a chance finding or a peculiarity of the specific dataset used. Essentially, the query arises: Is 0.72 statistically lower than 1? Should it be, it might lend support to Keynes's theory.

   The process of confirming or disproving economic theories based on sample evidence relies on a branch of statistical theory termed **statistical inference (hypothesis testing).** Throughout this book, we will delve into the actual mechanics of this inference process.

7. **Making Forecasts or Predictions:** If the selected model doesn't disprove the hypothesis or theory in consideration, we can employ it to foresee future values of the dependent variable, denoted as Y, based on known or expected future values of the predictor variable, represented as X. For instance, let's say we aim to predict the average consumption expenditure for the year 2006. The GDP figure for 2006 was 11319.4 billion dollars. By substituting this GDP value into the Eq. (1.3), we calculate:

$$\hat{Y}2006 = -299.5913 + 0.7218 \times 11319.4$$
$$= 7870.7516 \text{ billion dollars} \qquad (1.4)$$

   This implies an estimated average forecasted consumption expenditure of around 7870 billion dollars, considering the GDP value. However, the reported consumption expenditure for 2006 was 8044 billion dollars. This suggests that the estimated model (Equation I.3.3) underestimated the actual consumption expenditure by approximately 174 billion dollars. Describing it as a forecast error, this discrepancy represents about 1.5 percent of the actual GDP value for 2006.

Notes

When we thoroughly explore the linear regression model in upcoming chapters, we'll assess whether such an error is deemed "small" or "large." However, it's crucial to acknowledge that such forecast errors are unavoidable due to the statistical nature of our analysis.

Another use of the estimated model [Eq. (1.3)] comes into play when considering the potential impact of policy changes. For instance, if the president proposes an income tax reduction, what will be the effect on income, consumption expenditure, and eventually on employment?

If, in response to this policy change, investment expenditure increases, macroeconomic theory indicates that the change in income resulting from a dollar change in investment expenditure is defined by the income multiplier (M), given by:

$$M = \frac{1}{1 - MPC} \tag{1.5}$$

Using the MPC of 0.72 obtained from Eq. (1.5), this multiplier equates to approximately $M = 3.57$. Consequently, a one-dollar increase (or decrease) in investment will potentially yield more than a threefold increase (or decrease) in income, recognizing that the multiplier effect takes time to materialize.

The estimation of MPC plays a pivotal role in this calculation, as the multiplier hinges upon it. Regression models like Eq. (1.3) facilitate obtaining this MPC estimate, offering valuable insights for policy considerations. With knowledge of MPC, one can forecast the future trajectory of income, consumption expenditure, and employment following alterations in the government's fiscal policies.

8. **Applying the Model for Control or Policy Purposes:** In the final step, the hypothetical scenario where we possess the estimated consumption function as presented in Eq. (I.3.3). Additionally, assume the government holds the belief that consumer expenditure, around 8750 billion dollars (measured in 2000 dollars), will maintain the unemployment rate at its existing level of approximately 4.2 percent in the early part of 2006. The question arises: What income level is necessary to ensure the specified consumption expenditure target?

Notes

Should the results derived from the regression analysis, as indicated in Eq. (I.3.3), appear reasonable, basic arithmetic would demonstrate that:

$$8750 = -299.5913 + 0.7218(GDP_{2006}) \quad (1.6)$$

This calculation yields $X = 12537$, roughly. Therefore, an income threshold of about 12537 billion dollars, given an MPC of approximately 0.72, would generate an expenditure totaling around 8750 billion dollars.

These computations propose that an estimated model holds utility for control or policy purposes. Through a suitable blend of fiscal and monetary policy strategies, the government can manipulate the control variable X to achieve the desired level of the target variable Y.

1.5 Introduction to Econometric Softwares

Econometric software provides a gateway to the practical application of statistical methods in economics, offering powerful tools for data analysis, model estimation, and hypothesis testing. These software platforms serve as essential aids in implementing econometric techniques, allowing economists and researchers to manipulate data, perform complex calculations, and visualize results efficiently. There are various econometric softwares available for data analysis and interpretation such as:

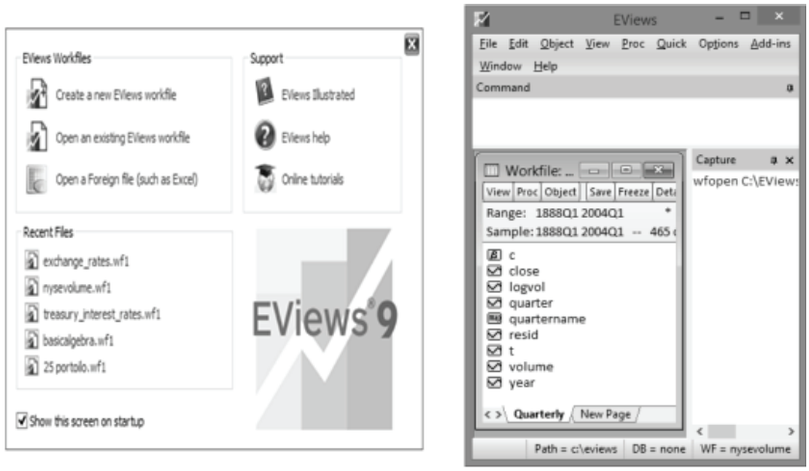
- EVIEWS
- GRETL
- R
- STATA
- EXCEL

1.5.1 *EViews*

EViews stands as a comprehensive econometric software package renowned for its robustness in time-series analysis and econometric modeling. It offers a user-friendly interface, facilitating data management, estimation of various econometric models, and insightful interpretation of results. EViews specializes in handling time-series data, enabling users to perform a wide

Notes

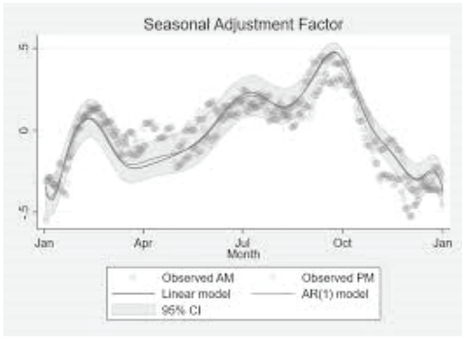
array of tasks, from basic descriptive statistics to advanced forecasting and panel data analysis, making it a favored choice among researchers, economists, and analysts dealing with time-series datasets.



(Source: https://www.eviews.com/illustrated/EViews%20Illustrated.pdf)

1.5.2 *STATA*

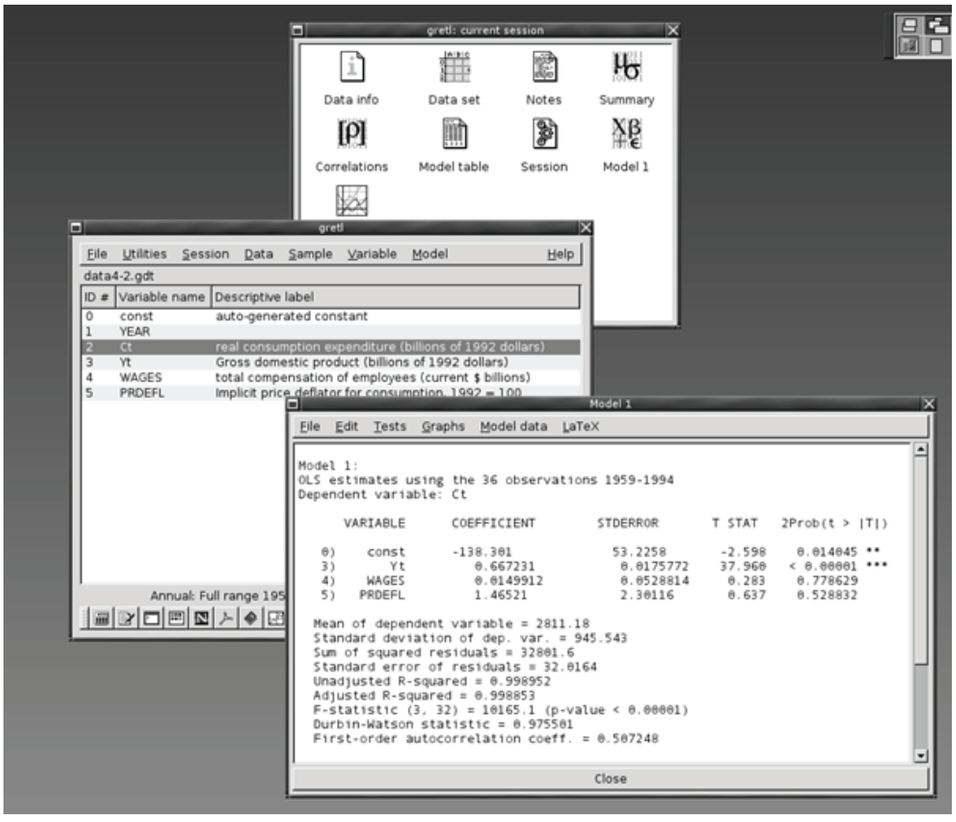
STATA is a powerful statistical software extensively used for data analysis, statistical modeling, and visualization in various fields, including economics. Renowned for its versatility, STATA supports a wide range of statistical techniques, from basic descriptive analysis to complex econometric modeling and panel data analysis. Its user-friendly interface, command-driven structure, and extensive capabilities for data management and manipulation make it a preferred choice for researchers, economists, and social scientists aiming to conduct sophisticated statistical analyses and generate publication-quality results.



Notes

1.5.3 *GRETL*

GRETL (Gnu Regression, Econometrics, and Time-series Library) is a user-friendly econometrics software designed for statistical analysis, econometric modeling, and time-series analysis. It offers an intuitive graphical interface combined with a powerful scripting language, enabling users to perform various statistical tasks, including data processing, regression analysis, time-series modeling, and forecasting. GRETL's simplicity and extensive range of econometric tools make it suitable for students, researchers, and practitioners seeking a user-friendly platform for econometric analysis and modeling.

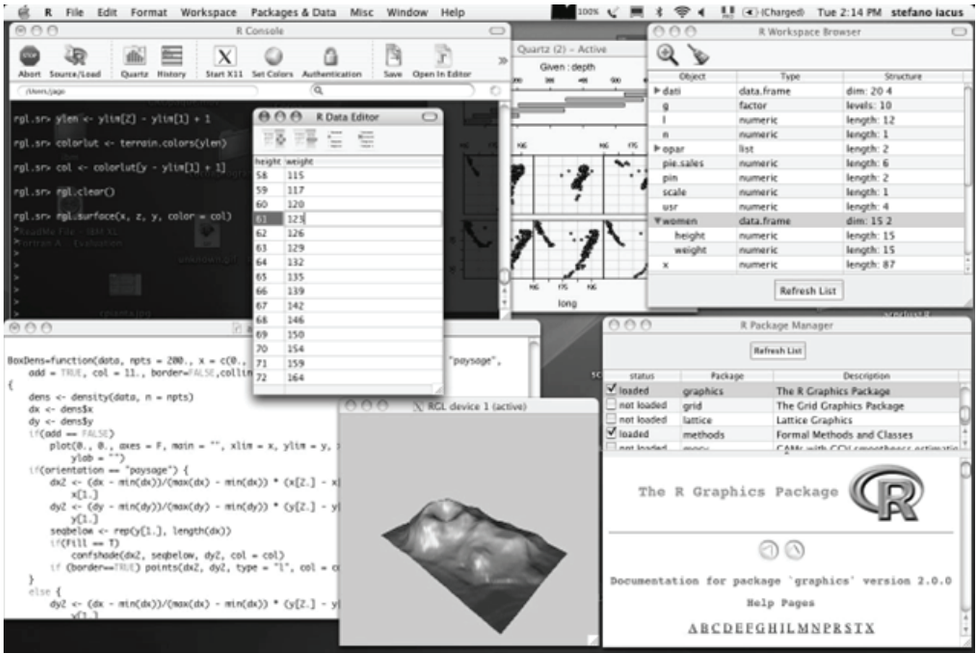


1.5.4 *R*

It seems like you've entered "R" as a standalone term. If you're referring to the statistical software called R, it's an open-source programming

Notes

language and software environment primarily used for statistical computing and graphics. It provides a wide range of statistical techniques, data analysis tools, and visualization capabilities, making it popular among researchers, statisticians, and analysts for conducting diverse statistical analyses and generating visual representations of data.



Throughout this book, we aim to enhance comprehension and engagement with diverse econometric concepts by integrating illustrative content and visual aids derived from EVIEWS software. By utilizing screenshots, graphical representations, and step-by-step demonstrations, we intend to offer a comprehensive and visual exploration of the intricacies within econometrics. These illustrations not only elucidate theoretical concepts but also provide practical insights into utilizing EVIEWS for model estimation, data analysis, and interpretation. Our objective is to facilitate a deeper understanding of econometric principles by presenting real-world applications and visual demonstrations using the EVIEWS platform, thereby fostering a richer learning experience for readers.

Notes

IN-TEXT QUESTIONS

1. What is the primary objective of specifying a mathematical model in econometrics?

(a) To simplify the understanding of complex economic theories

(b) To express economic theories in a mathematical format

(c) To eliminate the need for empirical verification

(d) To restrict the analysis to linear relationships only

2. In econometrics, what role does the "error term" play in a model?

(a) It represents the precision of the model's predictions

(b) It accounts for the variation in the dependent variable not explained by the model

(c) It is an indicator of the statistical significance of the model

(d) It represents the variability in the independent variables

3. Which statistical technique is commonly used in estimating the parameters of an econometric model?

(a) Descriptive statistics *(b)* Regression analysis

(c) Factor analysis *(d)* Cluster analysis

4. What does the term "hypothesis testing" involve in econometric modeling?

(a) Evaluating the model's predictions against actual economic data

(b) Assessing the economic theory's fit to the model

(c) Determining the statistical significance of model coefficients

(d) Checking for data outliers in the model

5. What purpose does the "forecasting" step serve in econometric modeling?

(a) It evaluates the accuracy of the model's predictions against historical data

(b) It estimates future values of the dependent variable using known values of the predictors

(c) It determines the strength of the relationship between variables

Notes

1.6 Summary

The topic of econometric modeling delves into the intricate process of applying statistical methods to economic theories and data. It involves constructing mathematical models to represent economic relationships, estimating model parameters, evaluating their significance, and utilizing these models for prediction and policy analysis.

Econometric modeling encompasses several key steps, starting with model specification, where economic theories are translated into mathematical equations. Estimation involves using statistical techniques like regression analysis to derive the parameters of the model. Hypothesis testing is crucial for evaluating the significance of these parameters, ensuring they align with the theoretical expectations.

Furthermore, forecasting plays a pivotal role in econometric modeling, allowing us to predict future values of economic variables based on historical data and model specifications. Illustrations and demonstrations using software like EVIEWS offer practical insights, aiding in the comprehension of these complex econometric concepts.

In essence, econometric modeling provides a systematic framework for analyzing economic phenomena, bridging economic theory and empirical analysis to derive meaningful insights for decision-making and policy formulation.

1.7 Answers to In-Text Questions

1. *(b)* To express economic theories in a mathematical format

2. *(b)* It accounts for the variation in the dependent variable not explained by the model

3. *(b)* Regression analysis

4. *(c)* Determining the statistical significance of model coefficients

5. *(b)* It estimates future values of the dependent variable using known values of the predictors

Notes

1.8 Self-Assessment Questions

1. Explain the significance of econometric modeling in understanding and analyzing economic phenomena, providing real-world examples to illustrate your points.
2. Describe the steps involved in model specification and their importance in econometric analysis. How does model specification affect the interpretation of results?
3. Discuss the role of hypothesis testing in econometric modeling. Explain how hypothesis testing aids in evaluating the significance of model coefficients and overall model fit.
4. Compare and contrast the applications of different econometric software (such as EVIEWS, STATA, R, or GRETL) in handling various types of economic data and conducting statistical analyses.
5. Illustrate the process of forecasting in econometrics, highlighting its challenges and the importance of forecast evaluation in assessing model accuracy.

1.9 References

- David A. Freedman, Robert Pisani, & Roger Purves. (2009). "Statistical Models: Theory and Practice", Cambridge University Press, ISBN-13: 978-0521743853.
- Ralph B. D. & Agostino. (1986). "Goodness of Fit Techniques", CRC Press, ISBN: 13: 978-0824784052.
- Alan Agresti. (2018). "An Introduction to Categorical Data Analysis", Wiley, ISBN: 13: 978-1119405269.

1.10 Suggested Readings

- David S. Moore, George P. McCabe, & Bruce A. Craig. (2017). "Introduction to the Practice of Statistics", W. H. Freeman, ISBN-13: 978-1464158933.
- John Fox. (2015). "Applied Regression Analysis and Generalized Linear Models", Sage Publications, ISBN-13: 978-1452205663.

LESSON

2

Regression Models: Assumption, Properties, Estimation and Hypothesis Testing

Dr. Ruhee Mittal
Assistant Professor
School of Open Learning
University of Delhi
Email-Id: ruhee.mittal@sol-du.ac.in

STRUCTURE

2.1 *Learning Objectives*
2.2 *Introduction*
2.3 *Model Specification*
2.4 *Estimation*
2.5 *Testing*
2.6 *Multicollinearity and Variable Inflation Factor (VIF)*
2.7 *Residual Analysis*
2.8 *Answers to In-Text Questions*
2.9 *Self-Assessment Questions*
2.10 *Summary*
2.11 *References*
2.12 *Suggested Readings*

2.1 Learning Objectives

- Define linear regression and understand the basic concepts.
- Explain the assumptions underlying linear regression.

Notes

- Understand the mathematical formulation of simple linear regression and multiple linear regression.
- Identify the assumptions of linear regression and understand their importance.
- Learn diagnostic techniques to assess the model's assumptions (e.g., residual analysis).
- Understand issues related to multicollinearity and how to address them using techniques like VIF.
- Dive deeper into residual analysis, understanding its importance in model validation.

2.2 Introduction

Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It is a supervised learning algorithm that aims to establish a linear relationship between the input features and the target variable. The primary goal of linear regression is to find the best-fitting linear equation that describes the relationship between the variables. Linear regression is a powerful and widely used tool for modelling and understanding relationships between variables. Its simplicity and interpretability make it a valuable technique for both prediction and inferential analysis in diverse domains.

Assumptions of Linear Regression: Linear regression makes several assumptions for its validity, including:

- **Linearity:** The relationship between variables is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** The variance of the errors is constant across all levels of the independent variables.
- **Normality of Residuals:** The residuals (errors) are normally distributed.
- **No Multicollinearity:** The independent variables are not highly correlated.

Applications: Linear regression finds applications in various fields, including:

- **Economics:** Modeling the relationship between variables like income and expenditure.

Notes

- **Finance:** Predicting stock prices based on various factors.
- **Biology:** Analyzing the impact of variables on biological processes.
- **Social Sciences:** Studying factors influencing human behavior.

2.3 Model Specification

2.3.1 *Simple Linear Regression*

A simple regression model is a statistical model that aims to describe the linear relationship between a single independent variable (predictor) and a dependent variable (response). The model assumes that this relationship can be represented by a straight line. The simple linear regression equation is typically expressed as:

Model Formulation:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- ***Y*:** The dependent variable (response variable) that we want to predict.
- ***X*:** The independent variable (predictor variable) that is used to predict *Y*.
- $\boldsymbol{\beta_0}$**:** Intercept or constant term. It represents the expected value of *Y* when *X* is zero.
- $\boldsymbol{\beta_1}$**:** Slope coefficient. It represents the change in the expected value of *Y* for a one-unit change in *X*, assuming a linear relationship.
- $\boldsymbol{\varepsilon}$**:** The error term. It captures the unobserved factors influencing *Y* that are not accounted for by the model. The goal is to minimize the sum of squared errors ($\sum \varepsilon^2$).

The objective of simple linear regression is to estimate the values of β_0 and β_1 that minimize the sum of squared differences between the observed values of *Y* and the values predicted by the regression line. This is often achieved using the method of least squares.

Key Components of a Simple Regression Model:

1. **Least Squares Estimation:** The process of finding the values of β_0 and β_1 that minimize the sum of squared differences between observed and predicted values.

Notes

2. **Intercept (β_0):** Represents the predicted value of Y when X is zero. It may or may not have a meaningful interpretation depending on the context.
3. **Slope (β_1):** Represents the change in the dependent variable for a one-unit change in the independent variable X. It indicates the direction and strength of the linear relationship.
4. **Residuals:** Differences between the observed values of Y and the values predicted by the regression line. Residual analysis is essential for assessing the model's fit and assumptions.
5. **Assumptions:** Similar to multiple regression, simple regression assumes linearity, independence of residuals, homoscedasticity (constant variance of residuals), and normality of residuals.

Applications of simple regression include predicting an outcome based on a single predictor variable, analyzing the strength and direction of the relationship between two variables, and making predictions or estimates for new observations.

While simple linear regression is straightforward and easy to interpret, it may not capture complex relationships in the data. For scenarios with more than one predictor, multiple linear regression is employed. Simple regression serves as a foundational concept for understanding the principles of linear modeling and is a valuable tool in various fields, including economics, finance, and social sciences.

2.3.2 *Multiple Linear Regression*

Multiple linear regression is an extension of simple linear regression, allowing for the modeling of relationships between a dependent variable and multiple independent variables. The model assumes that the relationship between the dependent variable and the independent variables is linear and can be expressed by the equation:

Model Formulation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

- ***Y*:** The dependent variable (response variable).
- **$X_1, X_2, \ldots, X_n$:** Multiple independent variables (predictors).
- **β_0:** Intercept or constant term.

Notes

- $\beta_1, \beta_2, \ldots, \beta_n$**:** Coefficients for the respective independent variables. They represent the change in the expected value of Y for a one-unit change in the corresponding X, assuming all other variables are held constant.
- ε**:** The error term.

The coefficients $\beta_1, \beta_2, \ldots, \beta_n$ are estimated using the method of least squares, which minimizes the sum of squared differences between the observed values of Y and the values predicted by the regression equation.

Key Components of a Multiple Linear Regression Model:

1. **Intercept (β_0):** Represents the expected value of Y when all predictor variables ($X_1, X_2, \ldots, X_n$) are zero. Interpretation can be meaningful or not, depending on the context of the data.
2. **Slope Coefficients ($\beta_1, \beta_2, \ldots, \beta_n$):** Represent the change in the expected value of Y for a one-unit change in the corresponding predictor, holding other predictors constant. Interpretation is specific to the units of the corresponding predictor variable.
3. **Multiple Predictors:** $X_1, X_2, \ldots, X_n$ are the independent variables contributing to the model. Each variable provides additional information to predict the dependent variable.
4. **Residuals:** Differences between the observed values of Y and the values predicted by the multiple regression equation. Residual analysis helps assess the model's fit and assumptions.
5. **Assumptions:** Multiple linear regression assumes linearity, independence of residuals, homoscedasticity (constant variance of residuals), and normality of residuals, similar to simple linear regression.

Applications of multiple linear regression include predicting outcomes based on several predictor variables, understanding the relationships between multiple factors and an outcome, and identifying the relative importance of different predictors.

Advantages of multiple linear regression include the ability to model complex relationships and account for the influence of multiple variables simultaneously. However, challenges may arise, such as multicollinearity (high correlation between predictors) and potential overfitting if the model becomes too complex.

Notes

In summary, multiple linear regression is a powerful statistical tool for modeling the relationships between a dependent variable and multiple independent variables, providing a versatile framework for analysis in various fields, including economics, social sciences, and natural sciences.

2.4 Estimation

Estimation in the context of linear regression involves finding the values for the coefficients that best fit the model to the observed data. The most common method for this purpose is the method of least squares. This method minimizes the sum of the squared differences between the observed values and the values predicted by the regression equation.

2.4.1 *Estimation in Simple Linear Regression*

For simple linear regression, where there is only one independent variable (predictor), the model is expressed as:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The coefficients β_0 and β_1 are estimated to minimize the sum of squared residuals:

Minimize $\sum_{i=1}^{n} = (Y_i - (\beta_0 + \beta_1 X_i))^2$

The estimates for β_0 and β_1 that minimize this sum are given by:

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}$$

where $\bar{X}$ and $\bar{Y}$ are the means of the independent and dependent variables, respectively.

2.4.2 *Estimation in Multiple Linear Regression*

For multiple linear regression, where there are multiple independent variables, the model is expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon$$

Notes

The estimates for the coefficients β_1 β_2, …, β_n that minimize the sum of squared residuals are obtained by solving a system of equations known as the normal equations. The matrix form of the normal equations is:

$$\mathbf{X^T Xb = X^T Y}$$

where:

- **X** is the matrix of independent variables,
- **Y** is the vector of dependent variable values,
- **b** is the vector of coefficient estimates.

The solution for b is given by:

$$\mathbf{b = (X^T X)^{-1} X^T Y}$$

This formula provides the estimates for β_1, β_2, …β_n, that minimize the sum of squared residuals.

In both cases, after obtaining the coefficient estimates, the linear regression model can be used to predict the dependent variable for new observations based on the values of the independent variables.

2.5 Testing

2.5.1 *Hypothesis Testing*

Hypothesis testing in linear regression, whether simple or multiple, involves assessing the significance of the regression coefficients and overall model fit. The two primary hypotheses commonly tested are related to the individual coefficients (slope parameters) and the overall significance of the model.

Hypothesis Testing for Simple Linear Regression:

1. **Testing Individual Coefficients (β_0 and β_1):**
   - **Null Hypotheses:**
     - $H_0 : \beta_0 = 0$ (The intercept is equal to zero)
     - $H_0 : \beta_1 = 0$ (The slope is equal to zero)
   - **Alternative Hypotheses:**
     - $H_1 : \beta_0 \neq 0$ or $H_1 : \beta_0 > 0$ or $H_1 : \beta_0 < 0$
     - $H_1 : \beta_1 \neq 0$ or $H_1 : \beta_0 = 0$ or $H_1 : \beta_1 < 0$

Notes

Test Statistic: The t-statistic is used for testing individual coefficients.

Decision Rule: Reject the null hypothesis if the p-value is less than the chosen significance level (e.g., 0.05).

2. **Testing Overall Model Significance (ANOVA):**
   - **Null Hypothesis:**
     - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (None of the coefficients are significant)
   - **Alternative Hypothesis:**
   - H_1 : At least one $\beta_i \neq 0$ (At least one coefficient is significant)

Test Statistic: F-statistic is used for testing overall model significance.

Decision Rule: Reject the null hypothesis if the p-value is less than the chosen significance level.

Hypothesis Testing for Multiple Linear Regression:

1. **Testing Individual Coefficients (β_0, β_1, …, β_n):**
   - **Null Hypotheses:**
     - $H_0 : \beta_0 = 0, \beta_1 = 0, \ldots, \beta_n = 0$ (Each coefficient is equal to zero)
   - **Alternative Hypotheses:**
     - H_1 : $\beta_i \neq 0$ for at least one i (At least one coefficient is not equal to zero)
   - **Test Statistic:** t-statistic is used for testing individual coefficients.
   - **Decision Rule:** Reject the null hypothesis for a particular coefficient if the p-value is less than the chosen significance level.
2. **Testing Overall Model Significance (ANOVA):**
   - **Null Hypothesis:**
     - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (None of the coefficients are significant)
   - **Alternative Hypothesis:**
     - H_1 : At least one $\beta_i \neq 0$ (At least one coefficient is significant)
   - **Test Statistic:** F-statistic is used for testing overall model significance.
   - **Decision Rule:** Reject the null hypothesis if the p-value is less than the chosen significance level.

Notes

Interpretation:

- **p-value:** In hypothesis testing, the p-value indicates the probability of observing the test statistic (or more extreme values) under the assumption that the null hypothesis is true. A small p-value (typically less than 0.05) leads to the rejection of the null hypothesis.
- **Significance Level (α):** The chosen level of significance, often denoted as α (e.g., 0.05), represents the threshold below which the null hypothesis is rejected.
- **Test Statistics (t-statistic, F-statistic):** These statistics measure how far the estimated coefficients are from what would be expected under the null hypothesis. Large values of these statistics contribute to the rejection of the null hypothesis.

In summary, hypothesis testing in linear regression involves assessing the statistical significance of individual coefficients and the overall model fit. Rejection of the null hypothesis suggests that at least one predictor variable is significantly related to the response variable.

2.5.2 *Overall Model Fit*

- **F-test:** The F-test in the context of linear regression is used to assess the overall significance of the model. It compares the fit of the estimated model with a model that has no independent variables (i.e., a model with only an intercept). The null hypothesis for the F-test is that all coefficients of the independent variables are zero, meaning the model has no explanatory power.

  The F-statistic is calculated as the ratio of two mean squared errors:

$$F = \frac{= SSR/k}{\frac{SSE}{n-k-1}}$$

Where:

- SSR is the sum of squared residuals when the model is estimated with the independent variables.
- SSE is the sum of squared residuals when the model is estimated without the independent variables (i.e., the null model with only an intercept).

Notes

- k is the number of independent variables in the model.
- n is the number of observations.

Under the null hypothesis (H_0), if the model has no explanatory power (all coefficients are zero), the F-statistic follows an F-distribution with k and $n-k-1$ degrees of freedom. If the F-statistic is significantly different from 1, it suggests that the model is statistically significant.

Steps for Conducting an F-test in Linear Regression:

1. **Formulate Hypotheses:**

   Null Hypothesis (H_0): All coefficients are zero (model has no explanatory power).

   Alternative Hypothesis (H_1): At least one coefficient is non-zero (model is significant).

2. **Calculate the F-statistic:**

   Obtain the Sum of Squared Residuals (SSR) and Sum of Squared Errors (SSE) from the regression output.

3. **Determine Degrees of Freedom:**

   - Degrees of freedom for the numerator is k (number of coefficients being tested).
   - Degrees of freedom for the denominator is $n-k-1$ (total sample size minus the number of coefficients and 1).

4. **Compare with Critical Value or P-value:**

   - Use the F-statistic to look up a critical value from an F-distribution table or compare it with a significance level.
   - Alternatively, obtain the p-value associated with the F-statistic. If the p-value is less than the chosen significance level (commonly 0.05), you reject the null hypothesis.

If the F-test is significant, it indicates that at least one independent variable contributes significantly to explaining the variability in the dependent variable.

The F-test is crucial for assessing the overall significance of a regression model. If the F-statistic is significant, it suggests that the model explains a significant amount of variance in the dependent variable, and at least one predictor variable is contributing to the model's explanatory power.

Notes

2.5.3 *Goodness of Fit*

- **R-squared (R^2):** R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a linear regression model. It ranges from 0 to 1, where a higher R^2 indicates a better fit of the model to the data.

In the context of linear regression, R^2 is calculated as follows:

$$R^2 = 1 - \frac{SSR}{SST}$$

Where:

- SSR is the sum of squared residuals (the sum of the squared differences between the observed and predicted values of the dependent variable).
- SST is the total sum of squares, which measures the total variance of the dependent variable.

In simpler terms, R^2 is the proportion of the total variance in the dependent variable that is explained by the independent variables. A value of 0 indicates that the model does not explain any variability, while a value of 1 indicates that the model explains all the variability.

Interpreting R^2:

- $R^2 = 0$: The model does not explain any variability in the dependent variable.
- $R^2 = 1$: The model perfectly explains the variability in the dependent variable.
- $0 < R^2 < 1$: Indicates the proportion of variability explained by the model. For example, an R^2 of 0.75 means that 75% of the variance in the dependent variable is explained by the independent variables.

Limitations of R^2:

- R^2 does not indicate whether the coefficients are statistically significant.
- It does not provide information about the goodness of fit for models with different numbers of independent variables.
- High R^2 does not necessarily mean a causal relationship.
- R^2 may increase with the addition of irrelevant variables (overfitting).

Notes

It's important to use R^2 in conjunction with other statistical measures and consider the context of the analysis to assess the overall performance of the linear regression model.

In Summary: The linear regression model, when approached through matrices, allows for a concise and efficient representation. Model specification involves defining the relationships between variables, estimation employs the least squares method through matrix operations, and testing involves hypothesis tests and goodness-of-fit measures. Careful consideration of model assumptions and thorough testing are critical for robust and meaningful results in linear regression analysis.

2.6 Multicollinearity and Variable Inflation Factor (VIF)

Multicollinearity occurs in a multiple regression model when two or more independent variables are highly correlated. This correlation can lead to issues in the model, such as unstable coefficient estimates and difficulties in interpreting the individual contribution of each variable. Multicollinearity doesn't impact the overall predictive power of the model, but it can affect the reliability of the coefficient estimates.

Consequences of Multicollinearity:

1. **Unstable Coefficient Estimates:** The coefficients become highly sensitive to small changes in the data.
2. **Loss of Precision:** Standard errors for the coefficients can become inflated, leading to wider confidence intervals.
3. **Difficulty in Interpretation:** It becomes challenging to interpret the individual contribution of each variable to the dependent variable.
4. **Increased Risk of Type II Errors:** The model may falsely conclude that a variable is not important when, in fact, it is.

Detecting Multicollinearity:

1. **Correlation Matrix:** Examining the correlation matrix between independent variables.
2. **Variance Inflation Factor (VIF):** A quantitative measure indicating the severity of multicollinearity.

Variable Inflation Factor (VIF): VIF is a metric that quantifies how much the variance of an estimated regression coefficient increases if the

Notes

predictors are correlated. It is calculated for each predictor in a regression model.

Formula for VIF:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination obtained by regressing the i-th predictor against all the other predictors.

Interpretation of VIF:

- A VIF of 1 indicates no multicollinearity.
- A VIF between 1 and 5 suggests moderate multicollinearity that may be a concern depending on the context.
- A VIF above 5 indicates a high level of multicollinearity, and values above 10 are often considered problematic.

Handling Multicollinearity:

1. **Remove Redundant Variables:** If two variables are highly correlated, consider removing one.
2. **Combine Variables:** Create new variables as combinations of the correlated ones.
3. **Use Regularization Techniques:** Techniques like Ridge Regression can mitigate the impact of multicollinearity.
4. **Increase Sample Size:** Increasing the sample size may help, but it's not always practical.
5. **Principal Component Analysis (PCA):** Transform the original variables into uncorrelated principal components.

   Principal Component Analysis (PCA) is a dimensionality reduction technique that is not inherently tied to linear regression but is often used as a preprocessing step, especially when dealing with multicollinearity among independent variables. PCA transforms the original variables into a new set of uncorrelated variables called principal components. In the context of linear regression, PCA can be used to address issues related to multicollinearity and reduce the dimensionality of the feature space.

Notes

Key Steps of PCA:

1. **Standardization:**
   - Standardize the data by centering and scaling the variables to have a mean of 0 and a standard deviation of 1. This step is crucial as it ensures that variables with larger scales do not dominate the principal components.
2. **Covariance Matrix Calculation:**
   - Calculate the covariance matrix of the standardized variables. The covariance matrix represents the pairwise covariances between all pairs of variables.
3. **Eigenvalue Decomposition:**
   - Perform eigenvalue decomposition on the covariance matrix. This results in eigenvalues and corresponding eigenvectors. Eigenvectors represent the directions of maximum variance, and eigenvalues indicate the magnitude of variance along those directions.
4. **Selection of Principal Components:**
   - Sort the eigenvalues in descending order. The principal components are then selected based on the eigenvalues. The eigenvectors corresponding to the highest eigenvalues capture the most variance in the data.
5. **Projection of Data:**
   - Project the original data onto the selected principal components. This is achieved by multiplying the original data matrix by the matrix of selected eigenvectors.
6. **Reducing Dimensionality:**
   - Choose a subset of the principal components that capture a sufficiently high percentage of the total variance. This subset is typically determined by setting a threshold, such as retaining components that explain 95% or 99% of the variance.

Integration with Linear Regression: After performing PCA, the new principal components can be used as the predictors in a linear regression model. The advantages include:

Notes

1. **Mitigating Multicollinearity:**
   - PCA helps address multicollinearity by transforming the original correlated predictors into a set of uncorrelated principal components.
2. **Reducing Dimensionality:**
   - By retaining a subset of principal components, PCA reduces the dimensionality of the feature space, making the model more interpretable and potentially improving computational efficiency.
3. **Retaining Most of the Variance:**
   - Even with a reduced number of principal components, the retained components still capture most of the variance in the original data.

Considerations and Limitations:

1. **Loss of Interpretability:**
   - While PCA addresses multicollinearity, it may lead to a loss of interpretability, as the principal components are combinations of the original variables.
2. **Assumption of Linearity:**
   - PCA assumes a linear relationship between variables, and its effectiveness may be limited if the relationships are highly nonlinear.
3. **Interpretation Challenges:**
   - Interpreting the coefficients in the linear regression model becomes more challenging when using principal components instead of the original variables.

   PCA can be a valuable tool in linear regression when dealing with multicollinearity and high-dimensional data. It allows for the creation of uncorrelated variables that can be used as predictors in a regression model, potentially improving model performance and interpretability.

Dealing with multicollinearity is crucial for obtaining reliable and interpretable results in linear regression models. VIF is a valuable tool for identifying and assessing the severity of multicollinearity in the predictor variables.

Notes

2.7 Residual Analysis

Residual analysis is a crucial step in assessing the validity and appropriateness of a linear regression model. Residuals are the differences between the observed values (actual outcomes) and the values predicted by the regression model. Analyzing these residuals helps to evaluate the model's assumptions, identify potential problems, and improve the model's overall performance.

Residual analysis is a critical step in validating the assumptions of a linear regression model. It provides insights into the model's performance, highlights areas for improvement, and ensures the reliability of the statistical inferences drawn from the regression analysis.

Key Components of Residual Analysis:

1. **Residuals Plot:**
   - A scatter plot of residuals against the predicted values is often used to check for patterns or trends. A random spread of residuals around zero is desirable, indicating that the model is capturing the underlying relationships adequately.
2. **Normality of Residuals:**
   - Checking whether the residuals follow a normal distribution is essential for certain statistical inferences. A Q-Q plot or a histogram of residuals can be employed for this purpose.
3. **Homoscedasticity:**
   - Homoscedasticity refers to the constant variance of residuals across all levels of the independent variables. Residuals should exhibit an equal spread across the predicted values in a residual plot.
4. **Independence of Residuals:**
   - Residuals should not show any patterns or trends over time or across observations. Autocorrelation in residuals may indicate a violation of independence assumptions.
5. **Outliers and Influential Points:**
   - Identifying outliers and influential points is crucial. Outliers are extreme data points that may disproportionately influence the

Notes

regression model. Cook's distance and leverage plots are tools for detecting influential points.

Interpreting Residual Plots:

1. **Random Residuals Scatter:**
   - A random scatter of residuals around zero suggests that the model is capturing the underlying relationships well.
2. **Residuals Distribution:**
   - A normal distribution of residuals is desirable for making valid statistical inferences.
3. **Homoscedastic Residuals:**
   - An equal spread of residuals across predicted values indicates homoscedasticity.
4. **No Clear Patterns in Residuals:**
   - Residuals should not show clear patterns, trends, or cycles, suggesting that the model adequately captures the underlying structure.

Actions Based on Residual Analysis:

1. **Model Refinement:**
   - If patterns or trends are observed in residuals, consider refining the model by adding relevant predictors or transforming variables.
2. **Outlier Treatment:**
   - Address outliers by understanding their impact on the model. It may involve excluding extreme observations or transforming variables.
3. **Addressing Non-Normality:**
   - If residuals deviate from a normal distribution, consider transformations or robust regression techniques.
4. **Handling Heteroscedasticity:**
   - If the spread of residuals is not constant, consider transforming variables or using weighted least squares regression.
5. **Influential Point Assessment:**
   - Assess the impact of influential points on the model and consider adjusting the analysis accordingly.

Notes

IN-TEXT QUESTIONS

1. The correlation coefficient is used to determine:

(a) A specific value of the y-variable given a specific value of the x-variable

(b) A specific value of the x-variable given a specific value of the y-variable

(c) The strength of the relationship between the x and y variables

(d) None of these

2. If there is a very strong correlation between two variables then the correlation coefficient must be

(a) Any value larger than 1

(b) Much smaller than 0, if the correlation is negative

(c) Much larger than 0, regardless of whether the correlation is negative or positive

(d) None of these alternatives is correct

3. In regression, the equation that describes how the response variable (y) is related to the explanatory variable (x) is:

(a) The correlation model

(b) The regression model

(c) Used to compute the correlation coefficient

(d) None of these alternatives is correct

4. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 16 male college students by using least squares regression. The following regression equation was obtained from this study: $\hat{Y} = -0.0127 + 0.0180x$

(a) Each beer consumed increases blood alcohol by 1.27%

(b) On average it takes 1.8 beers to increase blood alcohol content by 1%

(c) Each beer consumed increases blood alcohol by an average of amount of 1.8%

(d) Each beer consumed increases blood alcohol by exactly 0.018

Notes

5. SSE can never be

(a) Larger than SST

(b) Smaller than SST

(c) Equal to 1

(d) Equal to zero

6. Regression modeling is a statistical framework for developing a mathematical equation that describes how

(a) One explanatory and one or more response variables are related

(b) Several explanatory and several response variables response are related

(c) One response and one or more explanatory variables are related

(d) All of these are correct.

7. In regression analysis, the variable that is being predicted is the

(a) Response, or dependent, variable

(b) Independent variable

(c) Intervening variable

(d) Is usually x

8. In least squares regression, which of the following is not a required assumption about the error term ε?

(a) The expected value of the error term is one

(b) The variance of the error term is the same for all values of x

(c) The values of the error term are independent

(d) The error term is normally distributed

9. Larger values of r^2 (R^2) imply that the observations are more closely grouped about the

(a) Average value of the independent variables

(b) Average value of the dependent variable

(c) Least squares line

(d) Origin

Notes

10. In a regression analysis if $r^2 = 1$, then

(a) SSE must also be equal to one

(b) SSE must be equal to zero

(c) SSE can be any positive value

(d) SSE must be negative

11. The coefficient of correlation

(a) Is the square of the coefficient of determination

(b) Is the square root of the coefficient of determination

(c) Is the same as r-square

(d) Can never be negative

12. In regression analysis, the variable that is used to explain the change in the outcome of an experiment, or some natural process, is called

(a) The x-variable

(b) The independent variable

(c) The predictor variable

(d) The explanatory variable

(e) All of the above (a-d) are correct

(f) None of the above is correct

13. In the case of an algebraic model for a straight line, if a value for the x variable is specified, then

(a) The exact value of the response variable can be computed

(b) The computed response to the independent value will always give a minimal residual

(c) The computed value of y will always be the best estimate of the mean response

(d) None of these alternatives is correct

Notes

14. A residual plot:

(a) Displays residuals of the explanatory variable versus residuals of the response variable

(b) Displays residuals of the explanatory variable versus the response variable

(c) Displays explanatory variable versus residuals of the response variable

15. In a regression and correlation analysis if $r^2 = 1$, then

(a) SSE = SST *(b)* SSE = 1

(c) SSR = SSE *(d)* SSR = SST

16. If the coefficient of determination is a positive value, then the regression equation

(a) Must have a positive slope

(b) Must have a negative slope

(c) Could have either a positive or a negative slope

(d) Must have a positive y intercept

17. If two variables, x and y, have a very strong linear relationship, then

(a) There is evidence that x causes a change in y

(b) There is evidence that y causes a change in x

(c) There might not be any causal relationship between x and y

(d) None of these alternatives is correct

18. If the coefficient of determination is equal to 1, then the correlation coefficient

(a) Must also be equal to 1

(b) Can be either −1 or +1

(c) Can be any value between −1 to +1

(d) Must be −1

Notes

19. In regression analysis, if the independent variable is measured in kilograms, the dependent variable

(a) Must also be in kilograms

(b) Must be in some unit of weight

(c) Cannot be in kilograms

(d) Can be any units

20. If the correlation coefficient is a positive value, then the slope of the regression line

(a) Must also be positive

(b) Can be either negative or positive

(c) Can be zero

2.8 Answers to In-Text Questions

1. *(c)* The strength of the relationship between the x and y variables

2. *(b)* Much smaller than 0, if the correlation is negative

3. *(b)* The regression model

4. *(c)* Each beer consumed increases blood alcohol by an average of amount of 1.8%

5. *(a)* Larger than SST

6. *(c)* One response and one or more explanatory variables are related

7. *(a)* Response, or dependent, variable

8. *(a)* The expected value of the error term is one

9. *(c)* Least squares line

10. *(b)* SSE must be equal to zero

11. *(b)* Is the square root of the coefficient of determination

12. *(e)* All of the above (a-d) are correct

13. *(a)* The exact value of the response variable can be computed

Notes

14. *(c)* Displays explanatory variable versus residuals of the response variable

15. *(d)* SSR = SST

16. *(c)* Could have either a positive or a negative slope

17. *(c)* There might not be any causal relationship between x and y

18. *(b)* Can be either −1 or + 1

19. *(d)* Can be any units

20. *(a)* Must also be positive

2.9 Self-Assessment Questions

1. Explain the concept of multicollinearity in linear regression. How does multicollinearity affect the interpretation of regression coefficients, and what techniques can be used to address it?
2. Describe the steps involved in building a linear regression model. Include the key considerations for selecting predictor variables and assessing the model's performance.
3. Discuss the assumptions of linear regression. How might violations of these assumptions impact the reliability of the regression analysis? Provide examples for each assumption.
4. Take a dataset with two independent variables and one dependent variable. Walk through the process of conducting a multiple linear regression analysis. Include the interpretation of coefficients and the assessment of model fit.
5. What is the purpose of residual analysis in linear regression? Explain how you would interpret a residual plot and what patterns might indicate issues with the model. Provide examples.
6. Compare and contrast simple linear regression and multiple linear regression. When would you choose to use one over the other, and what are the advantages and limitations of each?
7. Examine the concept of regularization in the context of linear regression. How do techniques like Ridge Regression and Lasso Regression work, and when might you choose to apply them?

Notes

8. Discuss the importance of feature engineering in the context of linear regression. Provide examples of how feature engineering can enhance the predictive performance of a regression model.
9. Explain the significance of the R-squared statistic in linear regression. How would you interpret a high or low R-squared value, and what are its limitations as an evaluation metric?
10. Imagine encountering outliers in your dataset. How might outliers impact a linear regression model, and what strategies would you employ to handle or mitigate their influence?

2.10 Summary

Linear regression is a statistical modeling technique that seeks to establish a linear relationship between a dependent variable and one or more independent variables. The fundamental premise of linear regression is to fit a linear equation to the observed data points, capturing the underlying patterns and trends in the relationships between variables. In its simplest form, known as simple linear regression, the model involves a single independent variable predicting the dependent variable. However, in multiple linear regression, multiple predictors contribute to the model, enabling a more comprehensive analysis of the interplay between various factors. The linear regression equation comprises coefficients for each predictor, including an intercept term, with the model trained to minimize the sum of squared differences between observed and predicted values. The model's accuracy is often assessed using metrics like R-squared and Mean Squared Error. Assumptions, such as linearity, independence of residuals, homoscedasticity, and normality of residuals, underpin the reliability of linear regression analysis. Additionally, the diagnostic tool of residual analysis is employed to validate model assumptions, detect outliers, and ensure the robustness of the regression results. Linear regression finds applications across diverse fields, from economics and finance to biology and social sciences, providing a versatile tool for predictive modeling and inferential analysis. While linear regression offers simplicity and interpretability, its effectiveness hinges on the careful consideration of assumptions, appropriate model selection, and thoughtful interpretation of results. In summary, linear regression is a foundational statistical method

Notes

with broad applicability, providing a framework for understanding and predicting relationships between variables in a wide range of real-world scenarios.

2.11 References

- Gareth James, Daniela Witten, Trevor Hastie, & Robert Tibshirani, "Introduction to Statistical Learning" ISBN-13: 978-1461471370.
- Sanford Weisberg , "Applied Linear Regression" ISBN-13: 978-0471315649.
- Shayle R. Searle, "Matrix Algebra Useful for Statistics" ISBN-13: 978-0471034716.

2.12 Suggested Readings

- David S. Moore, George P. McCabe, & Bruce A. Craig. (2017). "Introduction to the Practice of Statistics", W. H. Freeman, ISBN-13: 978-1464158933.
- John Fox. (2015). "Applied Regression Analysis and Generalized Linear Models", Sage Publications, ISBN-13: 978-1452205663.

UNIT - II

LESSON

3

Violations of Classical Assumptions 1

Neha Verma
Assistant Professor
Department of Economics
Kirori Mal College
University of Delhi
Email-Id: nverma@kmc.du.ac.in

STRUCTURE

3.1 *Learning Objectives*
3.2 *Introduction*
3.3 *Multicollinearity*
3.4 *Heteroscedasticity*
3.5 *Summary*
3.6 *Answers to In-Text Questions*
3.7 *Self-Assessment Questions*
3.8 *References*
3.9 *Suggested Readings*

3.1 Learning Objectives

- Understand the nature of problems due to violation of the assumptions of no multicollinearity and homoscedasticity in the Classical Linear Regression Model (CLRM).
- Learn how to diagnose these issues through statistical tests.
- Understand the impact on the reliability and validity of inferences in case of violation of these assumptions.
- Examine the remedies available to alleviate these problems.

Notes

3.2 Introduction

The classical linear regression model is a method to estimate the parameters of a model by the method of least squares. The model makes some crucial assumptions like linearity of the parameters, nonstochastic regressors, zero conditional mean of errors, no heteroscedasticity (homoscedasticity), no autocorrelation, no multicollinearity and no specification error.

However, these assumptions are often violated in the real-life data. The researcher needs to exercise caution while making inferences based on the classical linear regression model if these assumptions are violated. This chapter discusses in detail the consequences of violating the assumptions of no multicollinearity and homoscedasticity in the dataset.

The chapter also discusses the various ways of detecting the violation of assumptions and the remedial measures to be adopted in specific cases.

3.3 Multicollinearity

3.3.1 *Nature of Multicollinearity*

If two or more variables in a classical linear regression model have an exact linear relationship between them, then it is a case of *perfect multicollinearity.* This violates the assumption of the CLRM. Consider a regression model with k explanatory variables, X_1, X_2, …, X_k (where $X_1 = 1$ for all the observations and therefore, can be understood as an intercept term), then an exact linear relationship between the explanatory variables would imply the following

$$\theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_k X_k = 0 \qquad (3.1)$$

where θ_1, $\theta_{2,\ldots,}$ θ_k are constants and all of them are not zero simultaneously.

In a broader sense, if the explanatory variables are intercorrelated (even if not perfectly), then it is considered to be a problem of imperfect (or near) multicollinearity. The equation (3.1) can be modified as follows to allow for a stochastic error term, u_i in the case of imperfect multicollinearity,

$$\theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_k X_k + u_i = 0 \qquad (3.2)$$

Notes

Consider the following example where X_2 refers to marks in statistics, X_3 refers to marks in econometrics and X_4 refers to marks in microeconomics for 5 students in a class.

Table 3.1: Example - Multicollinearity

X_2	X_3	X_4
30	60	62
43	86	84
23	46	50
50	100	93
21	42	42

In this hypothetical data, $X_3 = 2X_2$, therefore the marks in statistics and econometrics are perfectly collinear for these students with the correlation coefficient of unity. However, the variable X_4 is created from the variable X_3 by adding the following numbers 2, −2, 4, −7 and 0. The variables X_3 and X_4 are not perfectly collinear but have a high correlation coefficient of 0.9951. The inclusion of these variables in a regression model simultaneously can create a problem of multicollinearity.

Note that the concept of multicollinearity as defined only refers to the linear relationships among the explanatory variables. The regression model with a *nonlinear relationship* between the explanatory variables, as depicted in equation (3.3) does not strictly violate the assumption of no multicollinearity. However, these variables will be highly correlated and would make the estimation of parameters difficult.

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \qquad (3.3)$$

where Y refers to the dependent variables, βs are the coefficients and X_i, X_i^2 represent the explanatory variables. The two explanatory variables in this equation are functionally related in a nonlinear manner.

There are various sources of multicollinearity in data. Some of the factors causing multicollinearity are explained below:

(i) The inclusion of the same information twice can lead to a perfect correlation between the explanatory variables. For example, if the

Notes

weight of an object is measured in pounds in variable X_1 and kilograms in variable X_2, then both variables will be perfectly correlated.

(ii) *Falling into a dummy variable trap* by including all the categories of the dummy variable simultaneously in a single equation.

(iii) *Common trend* in the data can also make the explanatory variables highly collinear. For example, if the performance of students is regressed on school expenditure on instructional materials (X_2) and expenditure on athletics (X_3), then there is a possibility of high correlation between X_2 and X_3 because wealthier schools tend to spend more on both the aspects in comparison to poorer schools.

(iv) *An overdetermined model* can also violate the assumption of no multicollinearity as the number of explanatory variables would be more than the number of observations.

(v) *Model specification* with various polynomial terms of the explanatory variables can also make these variables highly collinear.

3.3.2 *Impact of Multicollinearity on Estimation and Inference*

If two or more variables are highly correlated in a dataset, then intuitively the estimator fails to distinguish between the individual effects of the explanatory variables. In applied econometrics, the regression analysis facilitates the separation of partial effects of each explanatory variable (X) upon the dependent variable. But with multicollinearity, the explanatory variables become indistinguishable and the partial effects of each X cannot be estimated.

Take the following regression model as an example,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (3.4)$$

where $X_3 = 2X_2$. The parameter β_2 estimates the rate of change in the dependent variable (Y) for a unit change in X_2, keeping X_3 constant. But as X_2 and X_3 are perfectly collinear, one of them cannot be kept constant while the other one is changing. Therefore, there is no way to estimate the individual effects of the explanatory variables on Y and there is no unique solution of the individual regression coefficients.

Notes

In case the explanatory variables are imperfectly correlated, the estimation of regression coefficients β_2 and β_3 is possible with the estimates being unbiased and consistent. Therefore, the estimators remain BLUE in the case of imperfect multicollinearity.

The problem arises in the inference of the statistical significance of these estimated coefficients as their standard errors increase due to multicollinearity making the estimates imprecise. The inflated standard errors of the coefficients lead to the following:

- Confidence intervals are very large making the estimates less reliable.
- The estimated *t*-statistics are very small making the variables statistically insignificant.
- Although the *t*-statistics of the explanatory variables are insignificant, the overall R^2 of the model can be very high which can give an incorrect measure for the goodness of fit.
- Any small change in the dataset can affect the OLS estimators and their standard errors by a large magnitude.

Mathematically, the variances and covariance of estimated coefficients ($\widehat{\beta_2}$ and $\widehat{\beta_3}$) for regression model in equation (2.4) are given by the following equations

$$var(\widehat{\beta_2}) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \tag{3.5}$$

$$var(\widehat{\beta_3}) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \tag{3.6}$$

$$cov(\widehat{\beta_2}, \widehat{\beta_3}) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \tag{3.7}$$

where r_{23} refers to the coefficient of correlation between X_2 and X_3 and σ^2 is the variance of *Y*. If there is perfect multicollinearity, then the coefficient r_{23} is unity and the variances and covariance are infinite. However, in the case of imperfect multicollinearity, a high value of r_{23} increases the values of variances and covariances. The speed at which these values increase with respect to the correlation coefficients of the explanatory variables is defined as the *variance-inflating factor* and denoted by:

Notes

$$VIF = \frac{1}{(1 - r_{23}^2)} \tag{3.8}$$

As the correlation coefficient (r_{23}) approaches 1, the VIF approaches infinity. Thus, as the extent of correlation keeps on increasing, the variance of the estimator increases and makes the inference of statistical significance difficult.

3.3.3 *Detection of Multicollinearity*

Multicollinearity refers to high correlation among the explanatory variables and is, therefore, a sample phenomenon. Moreover, the degree of correlation among the variables is the important factor while identifying and rectifying multicollinearity. Note that a small degree of multicollinearity may exist in every equation, it is the severe correlation among the variables which makes the inferences invalid. There is no unique statistical method to detect multicollinearity but there are a few rules of thumb to facilitate its detection. These rules are explained below:

1. **High R^2 and Insignificant t-statistics:** As the presence of multicollinearity inflates the standard error of the estimators, the t-statistics are decreased in magnitude, thereby making them statistically insignificant. If simultaneously, the R^2 is very high (in excess of 0.8), then it is a classic symptom of multicollinearity.
2. **High Correlation Among the Explanatory Variables:** Before running a regression model, it is advisable to check pair-wise correlations among the various independent variable along with the significance of the correlation coefficient. If these coefficients are high (in excess of 0.8), then there could be existence of multicollinearity.
3. **High Value of VIF:** The Variance Inflation Factor (VIF) discussed above can also help detect multicollinearity. If the calculated VIF is higher than 10, the variables are said to be highly correlated.

3.3.4 *Remedial Measures for Multicollinearity*

The imperfect (or near) multicollinearity does not affect the unbiasedness and consistency of estimators. The estimators still remain BLUE.

Notes

Therefore, one of the suggestions to deal with multicollinearity is to *'do* nothing'. Some degree of multicollinearity exists in every equation; therefore, the researcher can keep working with it. But if the degree of multicollinearity is high in the regression model, there are some remedial measures to deal with it. These measures are explained below:

1. **Dropping Variable(s):** One of the most common ways to deal with multicollinearity is to drop one or more of the variables from the regression equation which are showing high degree of correlation. For example, if consumption is regressed on income and wealth, then there is high possibility of multicollinearity as income and wealth will be collinear. In such a situation, taking only one of the variables (from income and wealth) can help deal with multicollinearity. However, note that dropping one or more variables may lead to *specification bias* or misspecification of the regression model. Therefore, if economic theory suggests that both income and wealth should be included in the equation, then the variable should not be dropped and some other remedial measure should be tried.

2. ***A priori* Information:** It refers to the information that the researcher has about the regression coefficients based on the previous empirical work or relevant theoretical constructs. Then, such information can be used to modify the regression equation and deal with multicollinearity. For example, let Y represent the output level which depends on two factors of production, labour (X_2) and capital (X_3). The regression equation is as below:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{3.9}$$

If the researcher has information that the production technology gives double return to capital in comparison to labour, that is, she $\beta_3 = 2\beta_2$ expects then the regression equation can be modified as

$$Y_i = \beta_1 + \beta_2 X_{2i} + 2\beta_2 X_{3i} + u_i \tag{3.10}$$

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{3.11}$$

where $X_i = X_{2i} + 2X_{3i}$. Once the estimated value of β_2 is calculated, the same for can also be obtained using the relationship $\beta_3 = 2\beta_2$.

Notes

3. **Increasing the Sample Size:** Another way to deal with multicollinearity is to combine increase the sample size of the dataset. This can rectify the problem of multicollinearity because it is primarily a sample phenomenon and increasing the sample size reduces the width of the confidence interval making the inferences more reliable. Recall the formula for the variance of an estimator in multiple regression model

$$var(\widehat{\beta_2}) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \quad (3.12)$$

As the sample size increases, the term $\sum x_{2i}^2$ will generally increase, thereby reducing the variance of the estimator and correcting for the effect of multicollinearity. This will able estimation of β_2 more precisely.

4. **Transforming the Data:** An appropriate transformation of the explanatory variables can help deal with multicollinearity. Two of such transformations discussed here are, *first differencing* and *ratio transformation.* Consider the following time-series regression model where Y is consumption level, X_2 refers to income level and X_3 refers to level of wealth at a particular time and let X_2 and X_3 are highly collinear.

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (3.13)$$

The same model at time (t−1) would be

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \quad (3.14)$$

Transforming the equation to the first difference form (subtracting equations (3.13) and (3.14)) would give the following:

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2,t-1}) + \beta_3 (X_{3t} - X_{3,t-1}) + (u_t - u_{t-1}) \quad (3.15)$$

Running the regression on equation (3.15) would reduce the severity of multicollinearity without affecting the causality because although income and wealth are highly collinear, there is no reason to believe that their difference would also be collinear.

Another solution is *ratio transformation*, in which the model in equation (3.13) is transformed as follows

$$\frac{Y_t}{X_{3t}} = \beta_1 \left(\frac{1}{X_{3t}}\right) + \beta_2 \left(\frac{X_{2t}}{X_{3t}}\right) + \beta_3 + \frac{u_t}{X_{3t}} \quad (3.16)$$

Notes

Running the regression on equation (3.16) would reduce the problem of multicollinearity.

Note that though such transformations reduce the collinearity between the explanatory variables, they may distort the assumptions of the error term and can make the disturbances serially correlated. Moreover, there is loss of one observation in the case of first difference transformation. Therefore, such remedies should be used with caution.

3.3.5 *Econometric Example: Multicollinearity*

Consider an example where consumption (Y) is regressed on income (X_1) and wealth (X_2) as in equation (3.9). The two explanatory variables are known to have a high correlation of 85 per cent in this hypothetical data. The estimation output of the regression using EViews is presented below.

Equation: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUMPTION
Method: Least Squares
Date: 12/12/23 Time: 00:21
Sample: 2001 2020
Included observations: 20

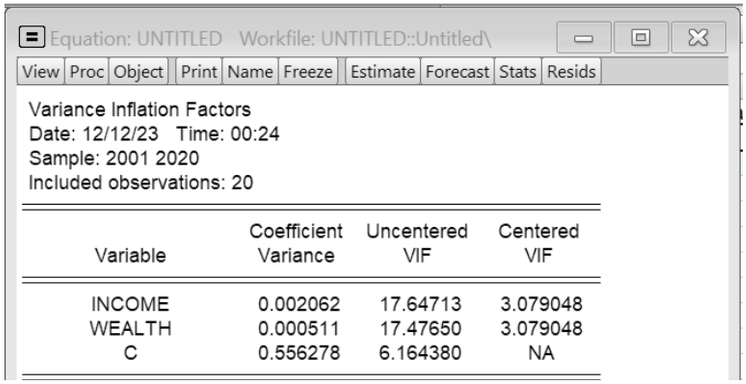
Variable	Coefficient	Std. Error	t-Statistic	Prob.
INCOME	0.916712	0.045409	20.18793	0.0000
WEALTH	-0.010539	0.022616	-0.465988	0.6471
C	0.939167	0.745841	1.259205	0.2250

R-squared	0.986122	Mean dependent var	23.55500
Adjusted R-squared	0.984489	S.D. dependent var	10.78691
S.E. of regression	1.343434	Akaike info criterion	3.565817
Sum squared resid	30.68186	Schwarz criterion	3.715176
Log likelihood	-32.65817	Hannan-Quinn criter.	3.594973
F-statistic	603.9696	Durbin-Watson stat	2.586886
Prob(F-statistic)	0.000000		

Figure 3.1: Estimation Output of Example: Multicollinearity

In the estimation output, note that the t-statistic for wealth is highly insignificant and the overall significance of the model (R^2) is 0.986, which is very high. These two observations are indications of the presence of multicollinearity in the data.

Notes


Equation: UNTITLED Workfile: UNTITLED::Untitled\

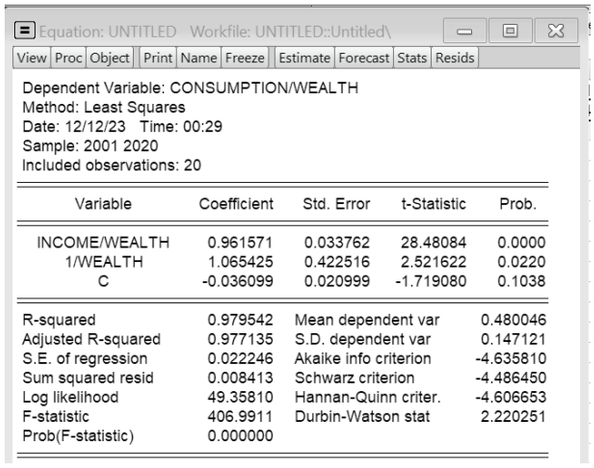
View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Variance Inflation Factors
Date: 12/12/23 Time: 00:24
Sample: 2001 2020
Included observations: 20

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
INCOME	0.002062	17.64713	3.079048
WEALTH	0.000511	17.47650	3.079048
C	0.556278	6.164380	NA

Figure 3.2: VIF

Moreover, the Variance-Inflation Factor (VIF) shown in the Figure 3.2 are higher than 10, thereby, suggesting the presence of multicollinearity in the data. To rectify the problem, one of the solutions is to do ratio transformation of the variables and then run the regression model. After taking the ratio of consumption to wealth as the dependent variable, along with dividing income and wealth with the variable wealth, the EViews estimation output is as follows:


Equation: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUMPTION/WEALTH
Method: Least Squares
Date: 12/12/23 Time: 00:29
Sample: 2001 2020
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
INCOME/WEALTH	0.961571	0.033762	28.48084	0.0000
1/WEALTH	1.065425	0.422516	2.521622	0.0220
C	-0.036099	0.020999	-1.719080	0.1038

R-squared	0.979542	Mean dependent var	0.480046
Adjusted R-squared	0.977135	S.D. dependent var	0.147121
S.E. of regression	0.022246	Akaike info criterion	-4.635810
Sum squared resid	0.008413	Schwarz criterion	-4.486450
Log likelihood	49.35810	Hannan-Quinn criter.	-4.606653
F-statistic	406.9911	Durbin-Watson stat	2.220251
Prob(F-statistic)	0.000000		

Figure 3.3: Ratio Transformation Regression

The t-statistics of the two independent variables are significant and the problem of multicollinearity is rectified.

Notes

IN-TEXT QUESTIONS

1. Which of the following situations does not indicate multicollinearity?

(a) High R^2 and insignificant t-ratios

(b) Wide confidence intervals

(c) Low standard errors

(d) Biased estimators

2. Which is not a way to rectify multicollinearity?

(a) Transformation by first difference

(b) Decreasing the sample size

(c) Dropping the variables

(d) Using a priori information

3.4 Heteroscedasticity

3.4.1 *Nature of Heteroscedasticity*

The assumption of a constant variance (σ^2) of each disturbance term, u_i in a classical linear regression model is termed as *homoscedasticity.* It implies equal (*homo)* spread (*scedasticity*) of the disturbance term as depicted in equation (3.17).

$$E(u_i^2) = \sigma^2 \qquad i = 1, 2, \ldots, n \tag{3.17}$$

The conditional variance of the dependent variable, Y_i (which is same as that of u_i) remains the same under homoscedasticity, regardless of the values taken by the explanatory variable, *X*. However, if this assumption is violated, then the conditional variance of *Y* may change with the change in the values of *X*. This is the case of *heteroscedasticity* and the variance of error terms is not constant under this condition (see equation (3.18)).

$$E(u_i^2) = \sigma_i^2 \qquad i = 1, 2, \ldots, n \tag{3.18}$$

In the Figure 3.4, as the values of *X* increase, the dispersion (spread) in the values of *Y* also increases. This is the case of heteroscedasticity.

Notes

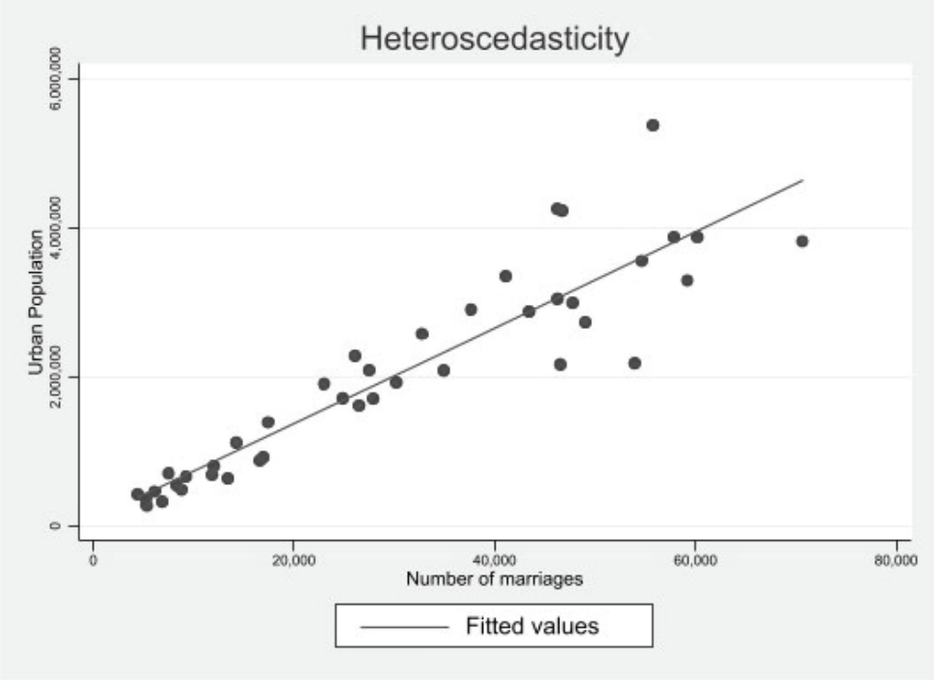


Figure 3.4: Example - Heteroscedasticity

There could be various reasons for getting heteroscedasticity in the data-set. Some of them are discussed below:

(i) **Error-learning Model:** These models suggest that as people learn, the errors arising in behaviour can reduce dramatically. For example, typing speed can improve with practice over time. Therefore, the σ_i^2 is expected to reduce as the value of independent variable increases (time in this example).

(ii) **Improved Data Collection Technique:** As the data collection techniques improve, the variance of errors can reduce over time.

(iii) **Presence of Outliers:** Outliers are the observations in the data which quite different (either small or large) from the rest of the observations of the data. Such observations can create heteroscedasticity in the data.

(iv) **Misspecification of Regression Model:** In case some variables are omitted from the model, it can lead to specification error. The error variance in such models may not be constant, thereby causing heteroscedasticity.

Notes

3.4.2 *Impact of Heteroscedasticity on Estimation and Inference*

Consider a two-variable regression model to understand the effect of heteroscedasticity on the OLS estimators

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{3.19}$$

In case of homoscedasticity, the OLS estimator for β_2 in this model (3.19) is given by the following

$$var(\widehat{\beta_2}) = \frac{\sigma^2}{\sum x_i^2} \tag{3.20}$$

However, in case of heteroscedasticity, σ^2 is not constant and variance of the estimator which changes to the following

$$var(\widehat{\beta_2}) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2} \tag{3.21}$$

The estimator of β_2 still remains linear, unbiased and consistent but it is no longer efficient, that is, it does not have the minimum variance in the class of unbiased estimators. The estimator of the variance of β_2 given by equation (3.21) is a biased estimator of the variance given by equation (3.20), that is, on an average it will underestimate or overestimate the latter with the bias being negative or positive respectively.

The consequences of using OLS estimators in the presence of heteroscedasticity are listed below:

- The estimated standard errors of the coefficients would be biased.
- As the estimators are no longer efficient, the estimated confidence intervals would be unnecessarily large.
- As a result, the estimated *t* and *F* statistics would be smaller than what is appropriate and therefore, would turn out to be insignificant.
- The usual testing procedures would provide misleading inferences in the presence of heteroscedasticity.

3.4.3 *Detection of Heteroscedasticity*

As in the case of multicollinearity, there are no set rules for detecting heteroscedasticity. There are certain rules of thumb which provide an

Notes

indication towards the existence of heteroscedasticity in the data. The rules are listed and explained below:

1. **Graphical Method:** The easiest way to detect heteroscedasticity is to run the regression model and examine the estimated residuals $\widehat{u_i^2}$ to check for any patterns visually. The estimated residuals are a good proxy for the true residuals in case the sample size is large and therefore, their graphical representation can help detect heteroscedasticity. The estimated residuals ($\widehat{u_i^2}$)can be plotted against the estimated values of the dependent variable $(\widehat{Y_i})$ from the regression model. If the resulting graph shows any set patterns (like linear, quadratic, parabolic), then it suggests that there is heteroscedasticity in the data and then the data should be accordingly transformed to take it into account. Consider the Figure 3.5 below.

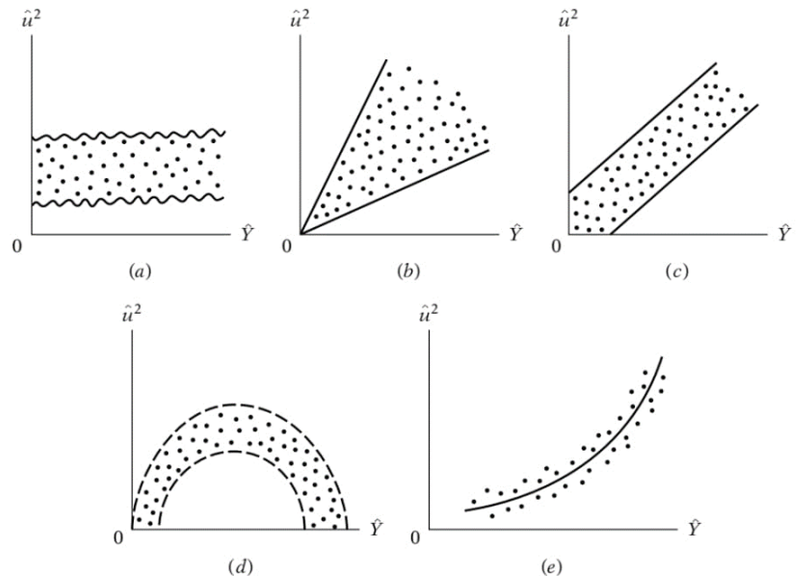


Figure 3.5: Detecting Heteroscedasticity: Graphical Method

In this figure, the panel (a) shows no exact relationship between estimated residuals and estimated values of Y, therefore, this case has no hetero-scedasticity. All the other panels (b, c, d, and e) show specific patterns and are the cases of heteroscedasticity.

2. **Park Test:** It is a formal method of detecting heteroscedasticity wherein σ_i^2 is some function of the explanatory variable, X_i. As the

true σ_i^2 is usually not known, the test suggests to use $\widehat{u_i^2}$ as a proxy for the same. The first step of the test is to run the OLS regression disregarding the heteroscedasticity and obtain the estimated residuals, $\widehat{u_i}$. Then, the following regression is run:

$$ln\widehat{u_i^2} = \alpha + \beta lnX_i + v_i \quad (3.22)$$

where v_i is the stochastic disturbance term.

In the regression shown in the equation (3.22), if the coefficient β turns out to be statistically significant, it would suggest the presence of heteroscedasticity in the data.

3. **Goldfeld-Quandt Test:** This test is applicable if it is assumed that the variance of errors under heteroscedasticity, σ_i^2 is positively related to one of the explanatory variables in the regression model. Following are the steps to conduct the test:

   - Arrange the data in the ascending order based on the values of the explanatory variable suspected to have caused heteroscedasticity, X_i.
   - Run two separate regressions, one for small values of X_i and one for the large values of X_i, after omitting the d observations for the middle.
   - Get the residual sum of squares for both the regressions as RSS_1 and RSS_2 for regression with smaller X_i and larger X_i respectively.
   - Calculate the following F ratio:

$$GQ = \frac{\frac{RSS_2}{df}}{\frac{RSS_1}{df}} \quad (3.23)$$

   If the residuals, u_i are assumed to be normally distributed and the assumption of homoscedasticity is valid, then GQ follows an F distribution with degrees of freedom for numerator and denominator as *(n-d-2k)/2*. Here k refers to the number of parameters to be estimated. For the two-variable case, k is equal to two.

   - The null hypothesis for the F-test is homoscedasticity and is tested against the alternative hypothesis of heteroscedasticity. If the estimated value of GQ is greater than the critical value of F at the chosen level of significance, we can reject the null of homoscedasticity in the data.

Notes

4. **White's General Heteroscedasticity Test:** It is the most popular test for heteroscedasticity as it does not rely on the normality assumption of the residuals and is easy to implement. Consider the following three-variable regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad (3.24)$$

The steps to be followed for White's test are:

- Estimate the equation (3.24) using OLS and estimate the residuals, $\widehat{u_i}$.
- Run the following auxiliary regression:

$$\widehat{u_i^2} = \alpha_0 + \alpha_1 X_{2i} + \alpha_2 X_{3i} + \alpha_3 X_{2i}^2 + \alpha_4 X_{3i}^2 + \alpha_5 X_{2i} X_{3i} + v_i \quad (3.25)$$

- Get the R^2 of this regression (overall significance of the model) and the sample size *n*.
- Estimate the following statistics

$$W = nR^2 \quad (3.26)$$

The statistic *W* will χ^2_{df} follow distribution, where *df* is equal to the number of regressors (excluding the constant term).

- If the value of the statistic, *W* is larger than the critical value of the chi-square distribution, then we reject the null hypothesis of homoscedasticity and the data exhibits heteroscedasticity.

3.4.4 *Remedial Measures for Heteroscedasticity*

The presence of heteroscedasticity does not affect the unbiasedness and consistency properties of the estimators, but they are no longer efficient, thereby making the inferences of hypothesis testing nonreliable. The remedial measures for the problem of heteroscedasticity depend on whether σ_i^2 is known or not known.

The remedies for heteroscedasticity are discussed below:

1. **Redefining the Variables:** The variance of the observations can be reduced by redefining the variables. For example, if the graphical method of detecting heteroscedasticity suggests that there are outliers in the data, the extreme values can be scaled down by taking logarithms of per capita values. Such transformation would reduce the variance of the residuals.

Notes

2. **Generalized Least Squares:** If σ_i^2 is known, the simplest way to correct for heteroscedasticity is by the mean of generalized least squares. Consider a two-variable regression model

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i \tag{3.27}$$

where $X_{0i} = 1$ for each i. As the heteroscedastic variances are known, the equation (3.27) is divided by σ_i to obtain the following:

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right) \tag{3.28}$$

which is the same as the following

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^* \tag{3.29}$$

where the starred variables represent the transformed variables. The parameters of the transformed model, β_1^* and β_2^* are different from the usual OLS estimators. The variance of transformed disturbance term is constant and therefore, the transformed model is homoscedastic. To prove this, notice the following feature of the transformed error term, u_i^*

$$var(u_i^*) = E(u_i^*)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2 \tag{3.30}$$

$$= \frac{1}{\sigma_i^2} E(u_i^2)$$

$$= \frac{1}{\sigma_i^2} \sigma_i^2$$

$$= 1$$

The above simplification assumes that σ_i^2 is known. The transformed model in equation (3.30) is, therefore, homoscedastic and satisfies the standard least-square assumptions.

3. **White's Heteroscedasticity-Consistent Standard Errors:** In case the σ_i^2 is not known, the researcher is suggested to use White's heteroscedasticity-consistent variances and standard errors which are available in almost every statistical package. These standard errors are also known as *robust standard errors.*

Notes

3.4.5 *Econometric Example: Heteroscedasticity*

Consider an example of consumption expenditure (Y) regressed on the income level (X) for 20 households and its estimation in EViews. The estimation output of the equation

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{3.31}$$

is as follows:

Equation: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CONSUMPTION
Method: Least Squares
Date: 12/11/23 Time: 23:47
Sample: 2001 2020
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.847052	0.703355	1.204302	0.2441
INCOME	0.899325	0.025309	35.53360	0.0000

R-squared	0.985944	Mean dependent var	23.55500
Adjusted R-squared	0.985164	S.D. dependent var	10.78691
S.E. of regression	1.313895	Akaike info criterion	3.478509
Sum squared resid	31.07377	Schwarz criterion	3.578082
Log likelihood	-32.78509	Hannan-Quinn criter.	3.497947
F-statistic	1262.637	Durbin-Watson stat	2.582686
Prob(F-statistic)	0.000000		

Figure 3.6: Estimation Output for Example: Heteroscedasticity

The residuals from the estimation are graphed to check the pattern of their variances with respect to the values of the independent variable.

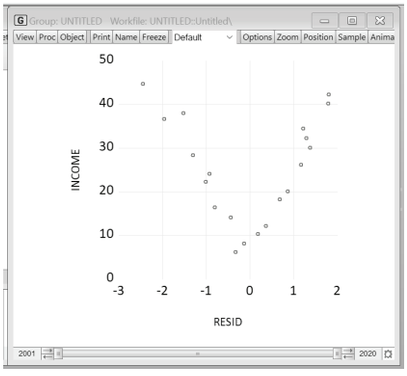


Figure 3.7: Representation of Residuals

Notes

In Figure 3.7, the residuals from the regression are plotted against the income level. For the high value of *X* (income level), the residuals are higher in magnitude and for the lower values of X, the residuals fall in magnitude. This suggests that the variances of residuals are not constant for different income levels. The detection of heteroscedasticity can also be done using White's Heteroscedasticity Test which has the following output in EViews.

Equation: UNTITLED Workfile: UNTITLED::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Heteroskedasticity Test: White
Null hypothesis: Homoskedasticity

F-statistic	116.7007	Prob. F(1,18)	0.0000
Obs*R-squared	17.32741	Prob. Chi-Square(1)	0.0000
Scaled explained SS	6.632037	Prob. Chi-Square(1)	0.0100

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 12/12/23 Time: 00:00
Sample: 2001 2020
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.256557	0.212171	-1.209199	0.2422
INCOME^2	0.002344	0.000217	10.80281	0.0000

R-squared	0.866370	Mean dependent var	1.553689
Adjusted R-squared	0.858947	S.D. dependent var	1.549642
S.E. of regression	0.582000	Akaike info criterion	1.849948
Sum squared resid	6.097039	Schwarz criterion	1.949521
Log likelihood	-16.49948	Hannan-Quinn criter.	1.869386
F-statistic	116.7007	Durbin-Watson stat	1.857106
Prob(F-statistic)	0.000000		

Figure 3.8: Output of White's Test

In Figure 3.8, the output of White's heteroscedasticity test is shown. The output of auxiliary regression yields the F-statistic of 116.7 which is higher than the critical value at the required degrees of freedom. Therefore, the null hypothesis of homoscedasticity is rejected and the data exhibits heteroscedasticity. The model can be re-estimated with White's heteroscedasticity-consistent standard errors to remove heteroscedasticity.

Notes

IN-TEXT QUESTIONS

3. In a multiple regression model if the error variances are not constant for each observation, we have the problem of

(a) Multicollinearity

(b) Heteroscedasticity

(c) Autocorrelation

(d) Specification Error

4. The problem of heteroscedasticity leads to

(a) Insignificant t-statistic

(b) Wide confidence intervals

(c) Biased standard errors

(d) All of these

3.5 Summary

The discussion on the violation of the assumptions of multicollinearity and heteroscedasticity can be summarized as follows:

1. The assumption of no multicollinearity implies that the two or more explanatory variables do not have high correlation coefficient.

2. In the presence of imperfect multicollinearity, the OLS estimates remain BLUE but the standard errors are inflated making the t-statistics insignificant. The overall level of significance (R^2) also gets inflated in this case.

3. In case of perfect multicollinearity, the estimates of the coefficients are indeterminate.

4. The detection of multicollinearity can be done by checking the correlation coefficient among the explanatory variables, checking if the value of R^2 is very high along with insignificant t-statistics and by checking if the value of the variance-inflation factor is higher than 10.

5. The problem of multicollinearity can be rectified by dropping the correlated variable, increasing the sample size, using a priori

Notes

information about the coefficients or by transforming the regression equation using first-difference or ratio transformation.

6. The assumption of no heteroscedasticity implies that the error variance is not constant, conditional on the explanatory variable.
7. In the presence of heteroscedasticity, the OLS estimators remain unbiased and consistent but they are no longer efficient. The estimated standard errors of the coefficients are biased which leads to unreliable hypothesis testing.
8. The detection of heteroscedasticity can be done by simple visual analysis of the residuals with respect to fitted values or the explanatory variables. Formal methods of detection include Park test, Goldfeld-Quandt Test and White's Heteroscedasticity test.
9. The remedial measures for heteroscedasticity include redefining the variables, using generalized least squares or using White's heteroscedasticity-consistent robust standard errors.

3.6 Answers to In-Text Questions

1. *(d)* Biased Estimators
2. *(b)* Decreasing the sample size
3. *(b)* Heteroscedasticity
4. *(d)* All of these.

3.7 Self-Assessment Questions

1. How the OLS estimators are affected in the presence of multicollinearity?
2. Explain the effects of heteroscedasticity on the estimates of the parameters in a multiple regression analysis.
3. How the use of generalized least squares correct for heteroscedasticity in the regression model?
4. Explain the concept of variance-inflation factor. How does it help in detection of multicollinearity in the model?

Notes

3.8 References

- Gujarati, N. Damodar. Basic Econometrics. New Delhi: McGraw Hill.
- Gujarati, N. Damodar. Econometrics by Examples. New Delhi: McGraw Hill.
- Christopher Dougherty. Introductory Econometrics. Oxford University Press.

3.9 Suggested Readings

- Maddala, G. S., & Lahiri, K. (1992). Introduction to econometrics (Vol. 2, p. 525). New York: Macmillan.
- Ramu, R. (2002). Introductory Econometrics with Applications (5th Edition). Thomson South-Western.

LESSON

4

Violations of Classical Assumptions 2

Neha Verma

Assistant Professor
Department of Economics
Kirori Mal College
University of Delhi
Email-Id: nverma@kmc.du.ac.in

STRUCTURE

4.1 *Learning Objectives*
4.2 *Introduction*
4.3 *Autocorrelation*
4.4 *Specification Errors*
4.5 *Summary*
4.6 *Answers to In-Text Questions*
4.7 *Self-Assessment Questions*
4.8 *References*
4.9 *Suggested Readings*

4.1 Learning Objectives

- Understanding the nature of problems due to violation of the assumptions of no autocorrelation and no specification errors in the Classical Linear Regression Model (CLRM).
- Learning how to diagnose these issues through statistical tests.
- Understanding the impact on the reliability and validity of inferences in case of violation of these assumptions.
- Examining the remedies available to alleviate these problems.

Notes

4.2 Introduction

The classical linear regression model is a method to estimate the parameters of a model by the method of least squares. The model makes some crucial assumptions like linearity of the parameters, nonstochastic regressors, zero conditional mean of errors, no heteroscedasticity (homoscedasticity), no autocorrelation, no multicollinearity and no specification error.

However, these assumptions are often violated in the real-life data. In the case of time-series data, where the observations are listed in a specific order based on time like days, months, years, etc, the successive observations may exhibit intercorrelations. Therefore, the assumption of no autocorrelation is usually violated in the case of time-series data and the error terms at different periods are serially correlated. If such correlation arises in cross-sectional data, then it is termed a *spatial correlation*.

The misspecification of the model also constitutes a violation of the classical linear regression model. Such misspecification includes incorrect independent variables, incorrect function form of the regression equation or incorrect form of stochastic error term.

The chapter also discusses the various ways of detecting the violation of assumptions and the remedial measures to be adopted in specific cases.

4.3 Autocorrelation

4.3.1 *Nature of Autocorrelation*

In simple words, autocorrelation refers to the correlation between the observations in a dataset ordered according to time or space. Consider the following linear regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{4.1}$$

The assumption of no autocorrelation means:

$$E(u_i u_j) = 0 \qquad i \neq j \tag{4.2}$$

Here i and j refer to the distinct time periods in the dataset (or different observations in cross-sectional data). It implies that the disturbance term related to any observation in the data is not influenced by the disturbance term relating to any other observation. For example, if we are regressing

Notes

the consumption expenditure of twenty distinct families on their income levels, the consumption level of a family is not affected by how the income level of another family affects their consumption level. In the case of time series data, consider the example where the movement of a stock price in period '1' is not affected by the price in period '0'. Usually, the real-life data violates the assumption of no autocorrelation (especially in the case of time series data). Symbolically, autocorrelation is depicted as follows

$$E(u_i u_j) \neq 0 \qquad i \neq j \tag{4.3}$$

Under autocorrelation, the consumption level for the households in the same neighbourhood can be correlated and the movement of a particular stock price may depend on past values its values. The equation (4.3) implies $Cov(u_i u_j) \neq 0,\ i \neq j$ where *Cov* refers to the covariance.

The following Figure 4.1 shows various patterns of autocorrelation and no autocorrelation in the series.

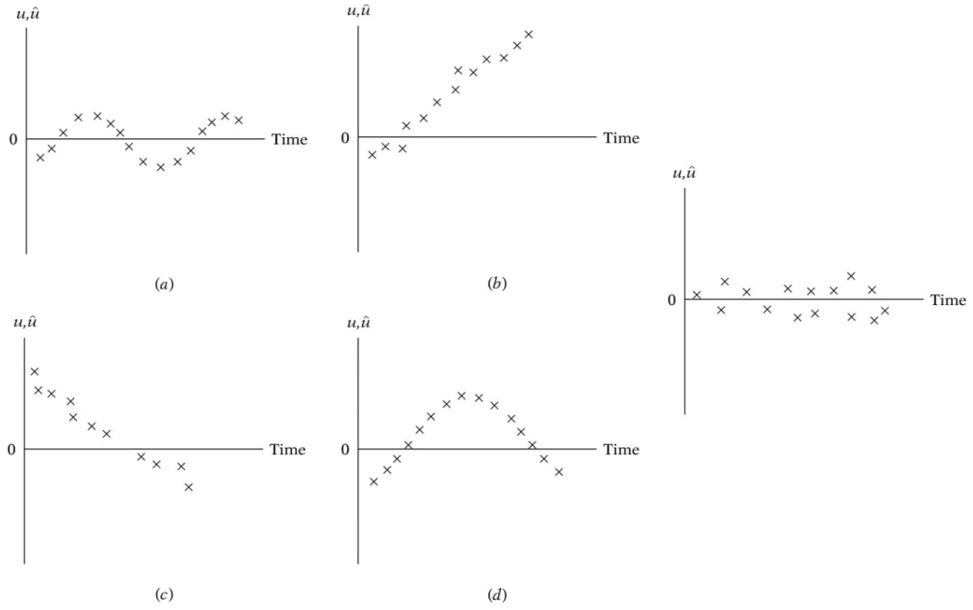


Figure 4.1: Patterns of Autocorrelation and No Autocorrelation

(Source: Gujarati (2004))

Panel (a) of Figure 4.1 shows a cyclical pattern in the residuals, panels (b) and (c) show the upward and downward trajectory of the residuals and panel (d) has a quadratic relation between the residuals. In panel (e) of

Notes

Figure 4.1, there is no pattern emerging from the graph of the residuals, therefore, it is the case of no autocorrelation.

Some of the possible cases of autocorrelation are discussed below:

1. **Model Misspecification:** The problem of autocorrelation can be caused by to incorrect functional form of the regression equation or the omission of certain variables in the model. Consider the following model as the correct specification of consumption expenditure

$$Consumption_i = \beta_1 + \beta_2 Income_i + \beta_3 Income_i^2 + u_i \quad (4.4)$$

But if the model in equation (4.5) is fitted instead

$$Consumption_i = \beta_1 + \beta_2 Income_i + v_i \quad (4.5)$$

Then we have a misspecified model because of the omission of the square of income from the equation. The disturbance term v_i actually includes $Income_i^2 + u_i$ in it and therefore, will catch the systematic effect of income level on consumption expenditure. Therefore, the disturbance term in equation (4.5) will violate the assumption of autocorrelation.

2. **Manipulation of Data:** In the time series analysis, the raw data is often averaged to smooth out the fluctuation in the data. For example, the quarterly data is converted to yearly data or monthly observations are converted to quarterly data. In such cases, a systematic pattern emerges in the disturbances due to the smoothening of data.

3. **Data Transformation:** Using of lagged values of the variables to construct a first difference model can lead to the problem of autocorrelation. Often the level form regression equation as in equation (4.6) is changed to a first difference equation as shown in equation (4.7) by subtracting the one period lagged values from the variables

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 X_t + u_t \\ Y_t - Y_{t-1} &= \beta_2 (X_t - X_{t-1}) + (u_t - u_{t-1}) \end{aligned} \quad (4.6)$$

$$\Delta Y_t = \beta_2 \Delta X_t + v_t \quad (4.7)$$

Note that subscript t is used to denote the time period. In equation (4.7), refers to the first difference operator and v_t refers to ($u_t - u_{t-1}$). The error terms in equation (4.6) satisfy the standard OLS assumptions,

Notes

but the errors in equation (4.7) are autocorrelated. The latter models are solved by dynamic regression models, which allow for lagged dependent and independent variables.

4. **Nonstationarity:** A time series is said to be stationary if its characteristics like mean, variance and covariance do not depend on time, that is, they are time-invariant. These characteristics do not change with time and the errors are not correlated. However, if the series are nonstationary, then its errors are autocorrelated.

4.3.2 *Impact of Autocorrelation on Estimation and Inference*

This section discusses the impact of autocorrelation in the disturbances on the OLS estimation and inference, when other assumptions of the classical linear regression model are not violated. Consider the following two-variable model of time-series data.

$$Y_t = \beta_1 + \beta_2 X_t + u_t \tag{4.8}$$

The mechanism that generates u_t has autocorrelation, that is, $E(u_t u_{t+s}) \neq 0\ (s \neq 0)$. A more specific form of assuming autocorrelation is as follows

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1 \tag{4.9}$$

where disturbance term in time t (u_t) is related to disturbance term in time period (t-1) (u_{t-1}) by *a coefficient of autocovariance* denoted by ρ (rho). In equation (4.9), the stochastic disturbance term is denoted by ε_t which satisfies the standard OLS assumptions. Therefore, has the following properties and is often referred to as a *white noise error term*

$$E(\varepsilon_t) = 0, Var(\varepsilon_t) = \sigma_\varepsilon^2 \text{ and } cov(\varepsilon_t, \varepsilon_{t+s}) = 0 \ \ s \neq 0 \tag{4.10}$$

The equation (4.9) postulates that the disturbance term in equation (4.8) at time period t is equal to rho times the disturbance term in time period (t-*1*) plus a purely random error term *(u_t)*. It is called first-order autoregressive scheme, denoted by AR(1). A similar scheme of autocorrelation could be second-order autoregressive scheme (AR(2)), which can be expressed as follows

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \varepsilon_t \tag{4.11}$$

In this lesson, we use AR(1) process because of its simplicity in application. The results will go through with higher degree of autocorrelation as well.

Notes

The coefficient of autocovariance in equation (4.9) is called first-order coefficient of autocorrelation of coefficient of autocorrelation at lag 1.

As the disturbance term in regression model (4.8) violates the assumption of no autocorrelation, the variance and covariance of error term with AR(1) autocorrelation scheme (equation 4.9) are denoted as follows

$$var(u_t) = E(u_t^2) = \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{4.12}$$

$$cov(u_t, u_{t+s}) = E(u_t u_{t-s}) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2} \tag{4.13}$$

$$cor(u_t, u_{t+s}) = \rho^s \tag{4.14}$$

where *var(u$_t$)* refers to the variance of the error term, $cov(u_t, u_{t+s})$ refers to the covariance between the error terms *s* periods apart and $cor(u_t, u_{t+s})$ refers to the correlation between the error terms *s* periods apart. The covariance and correlation coefficient follow the symmetry property, which implies that

$$cov(u_t, u_{t+s}) = cov(u_t, u_{t-s}) \; and \; cor(u_t, u_{t+s}) = cov(u_t, u_{t-s})$$

The degree of autocovariance is denoted by ρ, which lies between -1 and 1. If the value of ρ is equal to -1 or 1, the values of variance and covariance in equations (4.12) and (4.13) will not be defined. Therefore, $|\rho|<1$ and it implies that the value of covariance falls as the we go distant in the past (equation (4.12)). In other words, the series value in time period *t* is more correlated with the values in time period *t-1* than the values in *t-5*. The AR(1) process given in equation (4.9) is *stationary* if the mean, variance and covariance of u_t do not change overtime and it is the case when $|\rho|<1$.

The OLS estimators of βs and its variance for equation (4.8) get affected due to the AR(1) nature of the disturbance term. The effect of autocorrelation on the OLS estimators are listed below

1. *Estimated Coefficients* $(\widehat{\beta_2})$*:* The estimated coefficient of equation (4.8) remains linear, unbiased and consistent in the presence of autocorrelation of the error terms. But it is no longer BLUE as the $\hat{\beta}$ does not have minimum variance under the class of linear unbiased estimators. The estimator is not efficient.
2. *Biased Standard Error of* $\widehat{\beta_2}$ *:* The variance of the OLS estimator under AR(1) autocorrelation scheme has additional terms that depend on the coefficient of autocorrelation (ρ) and sample autocorrelations

Notes

between the values of the independent variable (X_t) at various lags. Therefore, the standard error of $\widehat{\beta_2}$ is biased and the degree of biasedness depends on the degree of autocorrelation in u_t. Serial correlation in errors causes the dependent variable to fluctuate in a way that the OLS estimation procedure attributes to the independent variable. It typically makes the OLS underestimate the standard errors of the coefficients.

3. *Tests of Significance:* As the standard error of the coefficients are no longer unbiased, the estimated *t-statistic* and *F-statistic* are incorrectly too high and therefore, the inferences are not valid and reliable about the statistical significance of the estimated regression coefficients.
4. *Confidence Interval:* As the estimated coefficient $\widehat{\beta_2}$ based on OLS is not BLUE, the confidence interval derived from its variance are incorrectly wider and therefore, the inference based on hypothesis testing is not reliable and valid.
5. R^2: The overall goodness of fit denoted by R^2 is overestimated in the presence of autocorrelation and use of OLS estimation technique.

4.3.3 *Detection of Autocorrelation*

The assumption of no autocorrelation implies that the disturbances across space or different time periods are not correlated with each other. The actual population disturbances denoted by u_t are not directly observables but the proxy for the same can be calculated from the sample data $\widehat{u_t}$. The detection of autocorrelation in the data can be done by both informal (graphical) and formal ways. The various methods of detection are explained in detail below:

1. **Graphical Method:** The visual inspection of $\widehat{u_t}$ (or $\widehat{u_t^2}$) graphed against time can provide the clues for autocorrelation in the population disturbance terms, u_t as the former is a good approximation of the latter. When we plot the sample residuals against time, it is called *time sequence plot*. Another option is to plot the standardized residuals against time. In order to obtain the standardized residuals, the value of each residual divided by the standard deviation of the residuals (standard error of the regression). The standardized

Notes

residuals are denoted by $(\widehat{u_t}/\hat{\sigma})$. The standardized residuals have zero mean and unit variance.

Figure 4.1 shows the different cases of time sequence plot where sample residuals are plotted against time. The cases (a), (b), (c), and (d) exhibit some pattern among the residuals and the panel (e) shows no pattern. Therefore, the latter case has no autocorrelation and the first four cases have autocorrelation in data.

2. **Runs Test:** This test attempts to discern the pattern in the residuals by noting their signs. It is a nonparametric test, also known as Geary test. The steps involved in Runs test are as follows:
   - The first step is not run a regression model using OLS and obtain the residuals, $\widehat{u_t}$.
   - The residuals are arranged according to time.
   - The number of runs formed by + and – signs are counted, where each run is defined as an uninterrupted sequence of one symbol (+ or -). The length of the run refers to the number of observations in it.
   - The number of runs in the data indicates the presence or absence of autocorrelation. If there are too many runs in the data, there is positive correlation among the residuals and if there are too few runs in the data, it is indicative of negative autocorrelation.
   - Let N_1 and N_2 are number of + symbols and – symbols respectively and R denote the number of runs. Then under the null hypothesis that the successive residuals are independent (no autocorrelation), and assuming $N_1 > 10$ and $N_2 > 10$, then the number of runs (R) is asymptotically normally distributed with following properties:

$$E(R) = \frac{2N_1N_2}{(N_1 + N_2)} + 1 \tag{4.15}$$

$$Var(R) = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N)^2(N-1)} \tag{4.16}$$

The Z statistic for conducting the Runs test can be calculated by subtracting E(R) from R and dividing by the standard deviation of R. The hypothesis testing can be done using the standard normal tables with the required level of significance.

Notes

3. **Durbin-Watson *d* Test:** The most commonly used test for autocorrelation is the Durbin-Watson *d* statistic test. The statistic is constructed using the sample residuals and is defined as follows

$$d = \frac{\sum_{t=2}^{t=n}(\widehat{u_t} - \widehat{u_{t-1}})^2}{\sum_{t=1}^{t=n}\widehat{u_t^2}} \tag{4.17}$$

The statistic *d* shown in equation (4.17) is constructed as the ratio of sum of squared differences in consecutive sample residuals to the residual sum of square (RSS). As the difference in the numerator is taken for successive residuals, one observation is lost and therefore, the number of independent observations is *n-1* instead of *n*. The statistic *d* follows certain assumptions as listed below:

- The regression model should include an intercept term for using this test.
- The independent variables, *Xs* should be fixed in repeated sampling, that is, nonstochastic.
- The population disturbances follow a first-order autoregressive scheme (AR(1)) as shown in equation (4.9).
- The error term u_t is assumed to be normally distributed.
- The regression models containing the lagged dependent variable (Y_{t-1}) as one of the independent variables cannot apply this test for checking the presence of autocorrelation.
- The *d* statistic does not control for missing observations; therefore, the data should not have any missing values.

The statistic *d* in equation (4.17) can be simplified as follows

$$d = \frac{\sum \widehat{u_t^2} + \sum \widehat{u_{t-1}^2} - 2\sum \widehat{u_t}\ \widehat{u_{t-1}}}{\sum \widehat{u_t^2}} \tag{4.18}$$

$$d \approx 2\left(1 - \frac{\sum \widehat{u_t}\ \widehat{u_{t-1}}}{\sum \widehat{u_t^2}}\right) \tag{4.19}$$

The equation (4.18) is changed to equation (4.19) by assuming $\sum \widehat{u_{t-1}^2} \approx \sum \widehat{u_t^2}$ as the two terms differ only by one observation. The sign refers to approximately. The *d* statistic in equation (4.19) can be written as

Notes

$$d \approx 2(1 - \hat{\rho}) \tag{4.20}$$

where $\hat{\rho} = \frac{\sum \widehat{u_t}\ \widehat{u_{t-1}}}{\sum \widehat{u_t^2}}$ is the estimated (from sample values) first-order coefficient of autocorrelation. As $-1 < \rho < 1$, therefore, the *d-statistic* in equation (4.20) lies between 0 and 4, that is $0 \leq d \leq 4$. Durbin and Watson have derived the lower bound d_L and upper limit d_U as the critical values to compare the calculated *d-statistic* against them and run the tests of significance. The critical values of d_L and d_U can be obtained from the Durbin-Watson tables based on the given sample size and the number of explanatory variables. The rule for the test of significance can be explained by the figure below

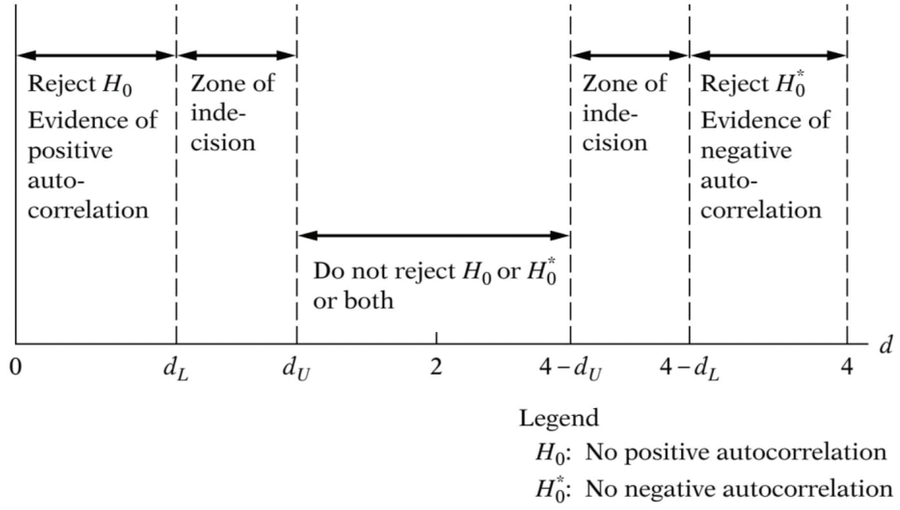


Figure 4.2: Rules of Test of Significance for *d-statistic*

(Source: Gujarati (2004))

The null hypothesis of the test is *no first-order autocorrelation in the disturbances* u_i. The decision to reject or not reject the null can fall in five different zones as shown in the Figure (4.2). If the estimated *d-statistic* lies between d_U and *4-*d_U, the null hypothesis is not rejected. If the *d-statistic* lies between 0 and d_L, the null hypothesis of no positive autocorrelation is rejected and the disturbances are positively correlated. On the other hand, if the estimated *d-statistic* lies between *4-*d_L and 4, the null hypothesis of negative autocorrelation is rejected. The zones between

Notes

d_L and d_U and between *4-*d_U and *4-*d_L are the zones of no decision.

When the *d-statistic* is equal to 0, there is perfect positive autocorrelation and when the *d-statistic* is equal to 4, there is perfect negative autocorrelation in the population disturbances.

4.3.4 *Remedial Measures for Autocorrelation*

If the data has autocorrelation based on the detection methods discussed in the previous sub-section, there are following remedial methods to deal with it.

1. **Correct Functional Form:** The omission of a key independent variable from the model can result in autocorrelation. Therefore, a simple way is to check the functional form of the regression model carefully and add a predictor variable which is probably missed. If adding such an independent variable can reduce or eliminate the autocorrelation in the model, then it should be added to the model.
2. **Add Lagged Dependent Variable:** One way to deal with autocorrelation in the data is to add the lagged dependent variable among the set of explanatory variables. This will control for the dynamics in the model and the autocorrelation can be corrected if the degree of serial correlation is low.
3. **Transformation of the Model:** This solution is similar to the one suggested for the presence of heteroscedasticity, wherein the model is appropriately transformed to nullify the effect the autocorrelation. The *Generalized Least Square Method* is discussed here.

   This method takes the autocorrelation among the residuals into account. The variables in the model are transformed as follows

$$Y_t' = Y_t - \hat{\rho} Y_{t-1} \tag{4.21}$$

$$X_t' = X_t - \hat{\rho} X_{t-1} \tag{4.22}$$

   The regression model run with the transformed variables as Y_t' and X_t' as the dependent and independent variables respectively correct for autocorrelation.
4. **First-Difference Form:** As the estimated ρ is subject to sampling errors, a simpler way is to run the regression model with first-differenced

Notes

dependent and independent variables. This measure is useful when the estimated *d-statistic* is small. The implicit assumption of this technique is that first difference of the errors is independent.

5. **White Robust Standard Errors:** The autocorrelation in the data does not affect the unbiasedness or consistency of the estimates but distorts the standard errors of the estimators which makes the inferences based on statistical tests invalid or unreliable. Therefore, one way to deal with this is to keep the estimated coefficients and adjust the standard errors by taking White robust standard errors. It is similar to the technique used for correcting heteroscedasticity as well.

4.3.5 *Econometric Example: Autocorrelation*

Consider an example where Consumer Price Index (CPI) is regressed on gross domestic product (GDP) and Index of Industrial Production (IIP) dataset for 2000Q1 – 2019Q4. The estimation output of regression is shown in Figure 4.3.

Equation: UNTITLED Workfile: MACRO_DATA::Macro_d...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CPI
Method: Least Squares
Date: 12/19/23 Time: 17:46
Sample: 2000Q1 2019Q4
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	20.51861	8.809158	2.329237	0.0225
GDP	0.000182	3.69E-05	4.943017	0.0000
IIP	0.174834	0.247664	0.705934	0.4824

R-squared	0.951554	Mean dependent var	103.8964
Adjusted R-squared	0.950296	S.D. dependent var	41.17561
S.E. of regression	9.179849	Akaike info criterion	7.308677
Sum squared resid	6488.761	Schwarz criterion	7.398003
Log likelihood	-289.3471	Hannan-Quinn criter.	7.344491
F-statistic	756.2054	Durbin-Watson stat	0.322118
Prob(F-statistic)	0.000000		

Figure 4.3: Estimation Output of Example: Autocorrelation

Notes

To check for autocorrelation in the dataset, the estimated residuals are plotted against the time. The resulting *time sequence plot* is shown below in Figure 4.4

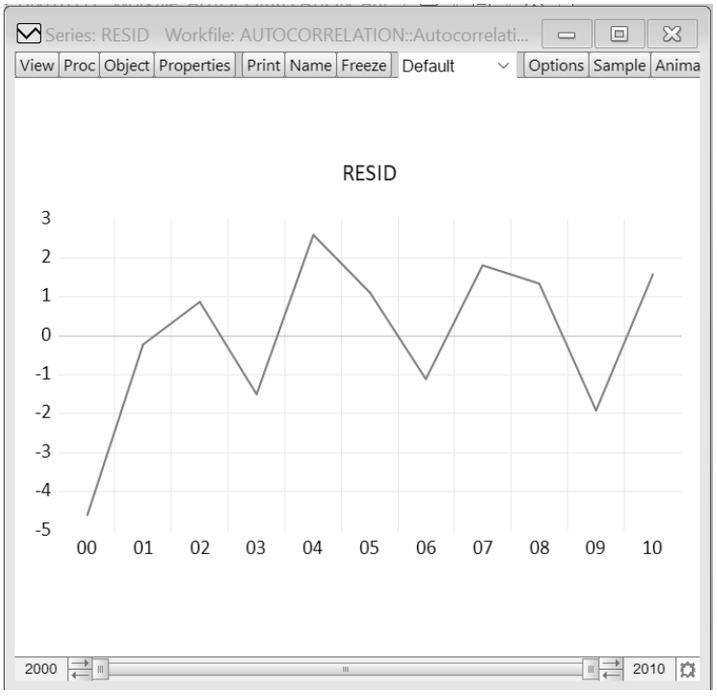


Figure 4.4: Graph of Residuals

The residuals graphed in Figure 4.4 seem to follow a pattern and are not random. The values are alternating between positive and negative values. To check formally, Durbin Watson test can be conducted. The estimation output in Figure 4.3 reports the Durbin-Watson Statistic (*d-statistic*) as 0.3221 at the bottom right corner of the table. The corresponding critical values for d_L and d_U can be obtained from the tables with number of independent variables *(k)* as 2 and number of observations as 80. The required critical values are 1.440 and 1.541 for d_L and d_U respectively. The *d-statistic of 0.3221*$<d_L$, therefore we can reject the null of no positive autocorrelation and there is evidence of autocorrelation in the disturbances.

We can deal with autocorrelation by adding a lagged dependent variable among the set of explanatory variables. In the figure below, the estimation output is presented when one-period lagged value of CPI is included in the regression

Notes

Equation: UNTITLED Workfile: MACRO_DATA::Macro_d...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: CPI
Method: Least Squares
Date: 12/19/23 Time: 17:52
Sample (adjusted): 2000Q2 2019Q4
Included observations: 79 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.639657	1.511600	-0.423165	0.6734
GDP	-2.66E-06	7.01E-06	-0.379494	0.7054
IIP	0.041406	0.040897	1.012449	0.3146
CPI(-1)	0.995253	0.018884	52.70303	0.0000

R-squared	0.998701	Mean dependent var	104.5383
Adjusted R-squared	0.998649	S.D. dependent var	41.03394
S.E. of regression	1.507964	Akaike info criterion	3.708703
Sum squared resid	170.5466	Schwarz criterion	3.828675
Log likelihood	-142.4938	Hannan-Quinn criter.	3.756768
F-statistic	19227.10	Durbin-Watson stat	1.692832
Prob(F-statistic)	0.000000		

Figure 4.5: Correction for Autocorrelation

In the estimation output above, the Durbin-Watson statistic is increased to 1.6928 and the corresponding critical values for three regressors and time period as 79 quarters are 1.416 and 1.568 for d_L and d_L respectively. The *d-statistic* lies between the d_U and *4-*d_U, therefore, the null hypothesis of autocorrelation is not rejected as per the decision rule of the test. Thus, the autocorrelation in the data is corrected by adding a lagged dependent variable. To further check the trajectory of residuals, the graph is presented in Figure 4.6.

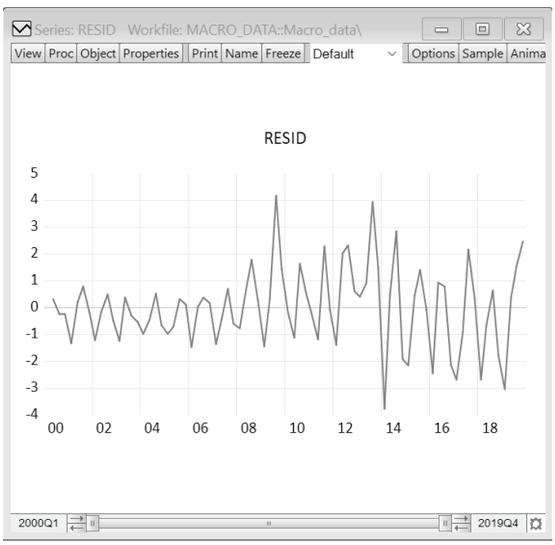


Figure 4.6: Correcting for Autocorrelation: Residuals

Notes

The graph of residuals in Figure 4.6 shows on specific pattern and the error seems to be random in nature. Therefore, there is no evidence of autocorrelation in the estimation.

IN-TEXT QUESTIONS

1. Which of the following situations does not usually cause autocorrelation?

(a) Model Misspecification

(b) Stationarity of the series

(c) Averaging of data

(d) Omission of relevant predictor variable

2. Which is not a way to detect autocorrelation?

(a) Runs test

(b) Durbin-Watson Statistic Test

(c) Variance Inflation Factor Test

(d) Time Sequence Plot

4.4 Specification Errors

4.4.1 *Nature of Specification Errors*

One of the assumptions of Classical Linear Regression Model (CLRM) is that the model is correctly specified. The violation of this assumption leads to a problem of *model specification bias* or *model specification error.* There are various types of specification errors can be encountered in regression modeling. Consider a model where the correct specification is as follows

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \quad (4.23)$$

Some of the possible specification errors in the model shown in equation (4.22) are discussed below

1. **Inclusion of Irrelevant Variable:** Suppose that instead of running the regression model in equation (4.23), the following model is run

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_i \quad (4.24)$$

Notes

In this model, the variable X_i^3 is not needed in the model. It is an irrelevant variable and adding it in the model is a specification error.

2. **Omission of a Relevant Variable:** Suppose that instead of model in equation (4.23), the following model is run

$$Y_i = \beta_1 + \beta_2 X_i + v_i \tag{4.25}$$

In this model, the variable X_i^2 is omitted from the model and the error term v_i will include $\beta_3 X_i^2 + u_i$ and therefore, the covariance between error and independent variable is no longer zero as assumed under CLRM. This would create a problem of endogeneity in the model.

3. **Wrong Functional Form:** Suppose that instead of the correct regression model in equation (4.23), the following model is run

$$lnY_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \tag{4.26}$$

In this model, instead of the level value of Y_i and natural logarithms of Y_i are regressed on the independent variables.

4. **Measurement Errors in the Data:** In case the dependent and independent variables are not measured correctly, that is, there are errors of measurement bias. Consider the following model with measurement errors:

$$Y_i' = \beta_1' + \beta_2' X_i' + \beta_3' X_i'^2 + u_i' \tag{4.27}$$

where $Y_t' = Y_i + \mu_i$ and $X_t' = X_i + \delta_i$, where μ and δ refer to the measurement errors independent and independent variables respectively.

4.4.2 *Impact of Specification Errors on Estimation and Inference*

The specification errors can occur in various ways as discussed in the previous subsection. The consequences of such errors on estimation and inference of regression coefficients in the cases of underfitting a model (omitting relevant variables) and overfitting a model (including irrelevant variables) are discussed below:

1. ***Underfitting a Model***

The underfitting of a model occurs in case a relevant variable if omitted from the regression equation. For example, suppose the true model is as follows:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \tag{4.28}$$

Notes

but instead, the following model is fit for running the regression analysis:

$$Y_i = \delta_1 + \delta_2 X_{2i} + v_i \quad (4.29)$$

In regression equation (4.29), the variable X_3 is missed as one of the explanatory variables. As the variable is relevant, its effect on the dependent variable (Y) is subsumed in the error term (v_i) and if X_2 and X_3 are correlated then the X_2 and v_i will also be correlated. This violates the assumption of zero covariance between explanatory variable and error term in the regression equation. Such a scenario creates the problem of endogeneity in the model.

Thus, in the case of underfitting the model, that is, omitting the relevant variable, the estimated coefficients (estimated values of δ_1 and δ_2) become biased and inconsistent. In case of no correlation between the included explanatory variable (X_2) and the omitted variable (X_3), δ_1 will have bias in estimation but δ_2 will have unbiased estimation.

The variance of the disturbances is also incorrectly estimated. Therefore, the usual tests of significance and confidence interval approach to test the null hypothesis are not valid and unreliable to assess the statistical significance of the estimated parameters. The forecasts made on the basis of incorrectly specified model are also not reliable.

2. *Overfitting a Model*

The overfitting of model implies the inclusion of an irrelevant variable in the model. Suppose the true model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \quad (4.30)$$

but instead, the following model is fit

$$Y_i = \delta_1 + \delta_2 X_{2i} + \delta_3 X_{3i} + v_i \quad (4.31)$$

In equation (4.31), an irrelevant variable (X_3) is added in the regression model. The OLS estimators of the parameters in equation (4.31) are all unbiased and consistent. It implies that $E(\delta_1) = \beta_1, E(\delta_2) = \beta_2$ and $E(\delta_3) = 0$. Moreover, the error variance is also correctly estimated in the incorrect model, thereby, validating the usual confidence interval and hypothesis-testing procedures.

Notes

The problem with overfitting of the model arises from the inefficiency of the estimated parameters. The variances of the estimated coefficients in regression equation (4.31) will be larger in magnitude than those of the parameters in the true model (equation (4.30)).

In case an irrelevant variable is incorrectly added in the model and it is highly correlated with one of the relevant variables in the model, then it can also lead to a problem of multicollinearity in the estimation.

4.4.3 *Detection of Specification Errors*

There are no set rules or techniques for detecting specification errors. However, the various ways of checking the specification errors in the model are discussed below:

1. *Detecting Overfitting of the Model:* A model has the problem of overfitting if irrelevant variables are included in the model. Let's take for example that in the following model, the variable X_2 is irrelevant.

$$Y_i = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + u_i \quad (4.32)$$

   Then, one simple to check for inclusion of irrelevant variable is to check for the statistical significance of β_2 and β_3. If the coefficient of X_2 is not statistically significant, there is a possibility that the variable is irrelevant for the regression model. An appropriate approach for this technique is to have *bottom-up approach* where the researcher starts with a smaller model and expands it by adding variables and checking their statistical significance. The variables which are not statistically significant can be dropped from the model (if economic theory warrants that).

2. *Detecting Underfitting of Model:* A model has a problem of underfitting if the relevant variables are omitted from the model. To start with it, after the initial estimation of the model, the broad features of the estimation output should be checked carefully for any unwarranted outcomes. The broad features like R^2, Durbin-Watson statistic, signs of estimated coefficients, *t-ratios* and *F-statistic.* If the results look encouraging then there may not be underfitting of model.

   Another way is to check the residuals of the estimation output. In case, the time sequence plot of residuals depicts distinct patterns, then there may be omission of a relevant variable or wrong functional form.

Notes

3. *R^2 and Adjusted R^2 Test:* The values of R^2 and *adjusted* R^2 can indicate about the specification errors. There is a tendency of R^2 to increase as more explanatory variables are added to the model, therefore, in a model with inclusion of irrelevant variables may lead to inflated R^2. To overcome this problem, the *adjusted* R^2 should be checked.

$$\overline{R^2} = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{4.33}$$

where n is the number of observations and k is the number of explanatory variables in the model. The adjusted R^2 takes into account the degrees of freedom of the model and therefore, penalizes for the addition of irrelevant variables. If the adjusted R^2 increases after the inclusion of a variable in the regression model, it is an indication that the variable is relevant and present in the true model.

4. *Information Criterion:* Another way of identifying the specification errors in the model is to check the information criterion for the different models. Some of the commonly used information criteria include Akaike Information Criterion (AIC) and Schwarz Information Criterion (SIC). The model with the lowest value of AIC or SIC has the best specification of the model.

5. *Ramsey's Regression Specification Error Test (RESET):* This is a formal test for specification error in the model. It is a test which checks whether the overall fit of the regression model can be improved by adding polynomials of the estimated dependent variable, $\widehat{Y}$. The steps of RESET test are as follows:

- Estimate the regression equation using OLS and save the fitted values of dependent value ($\widehat{Y_i}$).
- Create the polynomials of the fitted values like $\widehat{Y_i}^2$, $\widehat{Y_i}^3$, $\widehat{Y_i}^4$ and estimate the new regression equation with these polynomials added as individual regressors.
- Compare the fit of the two regression equations using the *F-test*, defined as follows

$$F = \frac{(R^2_{new} - R^2_{old})/(number\ of\ new\ regressors)}{(1 - R^2_{new})/(n - number\ of\ parameters\ in\ the\ new\ model)} \tag{4.34}$$

The *F-statistic* is checked against the critical values of *F-statistic* with the degree of freedom of the numerator is number of new

Notes

regressors and degree of freedom of the denominator as the difference between *n* and *number of parameters in the new model*. The null hypothesis of the test is that there is no misspecification. However, the null is rejected, there is presence of specification error but the type of specification error is not evident from this test.

4.4.4 *Remedial Measures for Specification Errors*

Fixing specification errors in the regression model is required for getting BLUE estimates and reliable tests for significance. The remedial measures to deal with specification bias are listed below:

1. In the case of omitted variable bias, the solution is to include the excluded variable or report the expected bias in the results for correct inference.
2. For fitting a correct model (or close to the true model), sequential specification search should be done. It involves sequentially dropping or including the variables and checking the broad features of the estimation output along with information criterion.

IN-TEXT QUESTIONS

3. Which of the following does not cause misspecification of the model?
   (a) Inclusion of irrelevant variables
   (b) Omission of relevant variables
   (c) Heteroscedasticity in the model
   (d) Wrong functional form
4. Which of the following test is used to test for specification error in the model?
   (a) Variance Inflation Factor Test
   (b) RESET test
   (c) Runs Test
   (d) All of these

Notes

4.5 Summary

The discussion on the violation of the assumptions of no autocorrelation and no specification error can be summarized as follows:

1. The concept of autocorrelation means that the population disturbances are correlated across time or across observations.
2. The presence of autocorrelation in the data does not affect the unbiasedness or consistency of the estimators, but the standard errors are biased, thereby making the tests of significance invalid and unreliable.
3. Autocorrelation can be detected using the time sequence plot of the residuals, Runs test or the Durbin-Watson Statistic test.
4. The remedial measures of autocorrelation include correcting the functional form of the model, using generalized least square estimation, taking first-difference model or adding the lagged dependent variable among the independent variables.
5. The types of specification error include underfitting the model, overfitting the model, measurement errors in the variables or wrong functional form. Such errors can make the estimators biased and distort the statistical tests of significance
6. The detection methods of specification errors include RESET test, information criterion method and checking the broad features of the estimation output.
7. To correct for specification errors, the solution is to include the excluded variable or using sequential specification search approach.

4.6 Answers to In-Text Questions

1. *(b)* Omission of relevant predictor variable
2. *(c)* Variance Inflation Factor Test
3. *(c)* Heteroscedasticity in the model
4. *(b)* RESET test

Notes

4.7 Self-Assessment Questions

1. How the OLS estimators are affected in the presence of autocorrelation?
2. Explain the effects of specification errors on the estimates of the parameters in a multiple regression analysis.
3. What are the different types of specification errors in the regression model?
4. How does the use of generalized least squares correct for autocorrelation in the regression model?
5. Explain the RESET test. How does it help in detection of specification errors in the model?

4.8 References

- Gujarati, N. Damodar. (2004). Basic Econometrics. New Delhi: McGraw Hill.
- Gujarati, N. Damodar. Econometrics by Examples. New Delhi: McGraw Hill.
- Christopher Dougherty. Introductory Econometrics. Oxford University Press.

4.9 Suggested Readings

- Maddala, G. S., & Lahiri, K. (1992). Introduction to Econometrics (Vol. 2, p. 525). New York: Macmillan.
- Ramu, R. (2002). Introductory Econometrics with Applications (5th Edition). Thomson South-Western.

UNIT - III

LESSON

5

Goodness of Fit

Dr. Tanu Kathuria
SDG Associate,
UNDP
Jammu & Kashmir, India
Email-Id: kathuriatanu@gmail.com

STRUCTURE

5.1 *Learning Objectives*
5.2 *Introduction*
5.3 *What is Goodness of Fit?*
5.4 *Test/Statistics Used for Goodness of Fit*
5.5 *R Square/*R^2
5.6 *Adjusted R Square/*Adjusted R^2
5.7 *Standard Error of the Model*
5.8 *Conceptual Understanding of AIC, BIC and SIC*
5.9 *Calculation and Comparison of AIC, BIC and SIC*
5.10 *Summary*
5.11 *Answers to In-Text Questions*
5.12 *Self-Assessment Questions*
5.13 *References*
5.14 *Suggested Readings*

5.1 Learning Objectives

- To determine how well the observed data fits with the model's expected value.
- Learning about different measures/tests/statistics used for goodness of fit.
- Understanding the usefulness of R-Square and Adjusted R-Square.
- Learning about the Standard Error of the model with respect to Goodness of Fit.
- Conceptual understanding, comparison and calculation of AIC, BIC and SIC.

Notes

5.2 Introduction

Goodness of fit is a fundamental concept in statistics that serves as a crucial tool for assessing how well a theoretical model aligns with observed data. This concept is central to various statistical analyses, aiding researchers and analysts in determining the validity of their models and hypotheses. The fundamental premise of goodness of fit involves comparing observed data with the expected values predicted by a statistical model. Whether applied to regression modeling, categorical data analysis, or the fitting of probability distributions, goodness of fit helps quantify the degree of agreement between theoretical expectations and real-world observations. Statistical tests, such as the Chi-Square test for categorical data or R-Square for regression models, play a pivotal role in this evaluation process. The significance of goodness of fit extends to model validation, hypothesis testing, and model selection, providing a systematic approach to ensure the robustness and reliability of statistical analyses in diverse fields.

5.3 What is Goodness of Fit?

Goodness of fit is a statistical concept that measures how well a model, or a distribution fits a set of observed data, implying that how well a model or set of predicted values fits the observed data. It quantifies the degree to which the observed data matches the values expected by the model. It can be used to test various hypotheses, such as whether the data follow a normal distribution, whether two samples come from the same population, or whether a regression model explains the variation in the data. In other words, it helps to evaluate whether the model provides a good representation of the real-world data.

5.4 Test/Statistics Used for Goodness of Fit

There are various statistical methods and tests used to assess goodness of fit, and the choice often depends on the type of data and the modelling approach. Some common methods include:

Notes

5.4.1 *Chi-Square Test*

The Chi-Square (χ^2) test is a statistical test used to determine whether there is a significant association between two categorical variables. It is a non-parametric test, meaning that it doesn't make assumptions about the distribution of the data. The Chi-Square test is commonly used in fields such as statistics, biology, social sciences, and market research.

There are two main types of Chi-Square tests:

Chi-Square Test for Independence: This test is used when you have two categorical variables, and you want to test whether they are independent or if there is an association between them.

The null hypothesis (H_0) assumes independence between the variables, and the alternative hypothesis (H_1) suggests dependence.

The formula for the test statistic (X^2) in the case of independence is:

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{(E_{ij})}$$

Where:

- O_{ij} is the observed frequency in the i-th row and j-th column of the contingency table.
- E_{ij} is the expected frequency in the i-th row and j-th column under the assumption of independence.

Chi-Square Test for Goodness of Fit: This test is used when you have one categorical variable, and you want to test whether the observed frequency distribution matches an expected distribution.

The null hypothesis (H_0) assumes that there is no significant difference between the observed and expected frequencies.

The formula for the test statistic (X^2) in the case of goodness of fit is:

$$X^2 = \sum \frac{(O_i - E_i)^2}{(E_i)}$$

Where:

- O_i is the observed frequency in the i-th category.
- E_i is the expected frequency in the i-th category.

Notes

In both cases, the test statistic (X^2) follows a Chi-Square distribution, and the p-value is used to determine the significance of the test. If the p-value is below a chosen significance level (commonly 0.05), the null hypothesis is rejected, suggesting that there is a significant association or difference between the variables. If the p-value is above the significance level, the null hypothesis is not rejected.

5.4.2 *Kolmogorov-Smirnov Test*

The Kolmogorov-Smirnov (KS) test is a non-parametric test used to determine whether a sample comes from a specific probability distribution. It is particularly useful for assessing the goodness-of-fit of a sample distribution to a theoretical distribution (e.g., normal, exponential).

The test is based on the Cumulative Distribution Function (CDF) of the observed sample and the CDF of the theoretical distribution. The KS statistic, denoted as D, represents the maximum vertical deviation between these two cumulative distribution functions. The null hypothesis (H_0) of the KS test is that the sample is drawn from the specified distribution.

The formula for the KS statistic is:

$$D = \max(\| F_0(x) - F_e(x) \|)$$

Where:

- $F_0(x)$ is the empirical (observed) cumulative distribution function of the sample.
- $F_e(x)$ is the cumulative distribution function of the theoretical distribution.
- The maximum is taken over all possible values of x in the sample.

The critical values of the KS statistic depend on the sample size and the chosen significance level. For a given significance level (e.g., 0.05), if the calculated KS statistic exceeds the critical value, the null hypothesis is rejected, suggesting that the sample does not come from the specified distribution.

It's important to note that the KS test is sensitive to differences in both location and shape of the distributions. While it can be used for any continuous distribution, it is most applied to test the goodness-of-fit for the normal distribution.

Notes

In some statistical software or programming languages, you may find variations of the KS test, such as the Lilliefors test, which is a modification of the KS test specifically designed for small sample sizes.

5.4.3 *Residual Analysis*

In regression analysis, residuals (the differences between observed and predicted values) are examined to check for patterns or trends. Analysing residuals helps assess how well the model fits the data and whether the assumptions of the regression model are met. A good fit is often indicated by random and evenly distributed residuals.

Residual analysis is an iterative process. If issues are identified, it may be necessary to modify the model or transform variables to address problems. It's important to note that no model is perfect, and residual analysis helps identify areas for improvement and assess the model's reliability.

Here are key aspects of residual analysis:

Residual Plots: Scatter plots of residuals against the predicted values or the independent variables are commonly used. These plots can reveal patterns or trends that may indicate issues with the model.

Ideally, the residuals should be randomly distributed around zero, with no clear patterns. Patterns, such as curves or systematic deviations, may suggest nonlinearity, heteroscedasticity, or omitted variable bias.

Normality of Residuals: Checking the normality of residuals is important, especially for linear regression models. Normality of residuals is often assumed for valid hypothesis testing and confidence interval estimation.

Q-Q plots (quantile-quantile plots) or histogram of residuals can be used to visually assess normality.

Homoscedasticity: It means that the variance of the residuals is constant across all levels of the independent variables. Heteroscedasticity (varying variance) can be problematic.

Residual plots against predicted values or independent variables can help identify heteroscedasticity.

Independence of Residuals: Residuals should be independent of each other. Autocorrelation or patterns in the residuals over time may indicate violations of independence.

Notes

Time series plots of residuals or Autocorrelation Function (ACF) plots can be used for detecting autocorrelation.

Outliers and Influential Points: Identify any influential data points or outliers that might strongly affect the model. Outliers can disproportionately influence regression coefficients.

Cook's distance and leverage statistics are measures used to identify influential points.

Residual Sum of Squares (RSS): The RSS is a measure of how well the model fits the data. Lower RSS indicates a better fit.

5.4.4 *Coefficient of Determination (R-squared)*

In regression models, R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables. A higher R-squared value suggests a better fit.

5.4.5 *Likelihood Ratio Test*

Likelihood Ratio Tests (LRTs) are statistical tests used in hypothesis testing and model comparison, particularly in the context of maximum likelihood estimation. These tests help assess whether adding or removing parameters from a statistical model significantly improves or degrades its fit to the data. These tests compare the likelihood of the data under the fitted model with the likelihood under a null model (no relationship).

Goodness of fit is important for ensuring the validity and reliability of statistical models and analyses. A good fit means that the model or the distribution captures the essential features of the data and can be used for prediction or inference. The assessment of goodness of fit is crucial in determining whether a statistical model is appropriate for the given data.

If the goodness of fit is poor, it may indicate that the model needs improvement or that a different model should be considered. A poor fit means that the model or the distribution is not suitable for the data and may lead to erroneous or misleading conclusions.

On the other hand, a good fit suggests that the model is a reasonable representation of the underlying data patterns.

Notes

IN-TEXT QUESTIONS

1. Which of the following statements best describes the concept of goodness of fit in statistics?

(a) It measures how well the observed data matches the expected data in a statistical model.

(b) It assesses the variability of the dependent variable.

(c) It evaluates the correlation between two variables.

(d) It measures the central tendency of a dataset.

2. Which statistical test is commonly used for assessing goodness of fit in categorical data?

(a) T-test *(b)* ANOVA

(c) Chi-square test *(d)* Pearson correlation

3. What does the residual analysis involve in the context of goodness of fit?

(a) Examining the difference between observed and expected values.

(b) Calculating the mean of the dataset.

(c) Assessing the correlation coefficient.

(d) Analyzing the variance of the dependent variable.

5.5 R Square/R^2

5.5.1 *R-squared*

(R^2) is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model. It is also known as the coefficient of determination. R-squared is a key output in regression analysis and provides insight into the goodness of fit of the model.

Notes

Here's how R-squared is calculated:

$$R^2 = 1 - \frac{Sum\ of\ Squares\ of\ Residuals}{Total\ Sum\ of\ Squares}$$

Sum of Squares of Residuals (SSR): This represents the sum of the squared differences between the observed values and the values predicted by the model.

Total Sum of Squares (SST): This represents the sum of the squared differences between the observed values and the mean of the dependent variable.

The formula essentially compares how well the model performs compared to a simple model that uses the mean of the dependent variable to make predictions. If the model's predictions are completely off and no better than predicting the mean, R-squared would be close to 0. If the model perfectly predicts the observed values, R-squared would be 1.

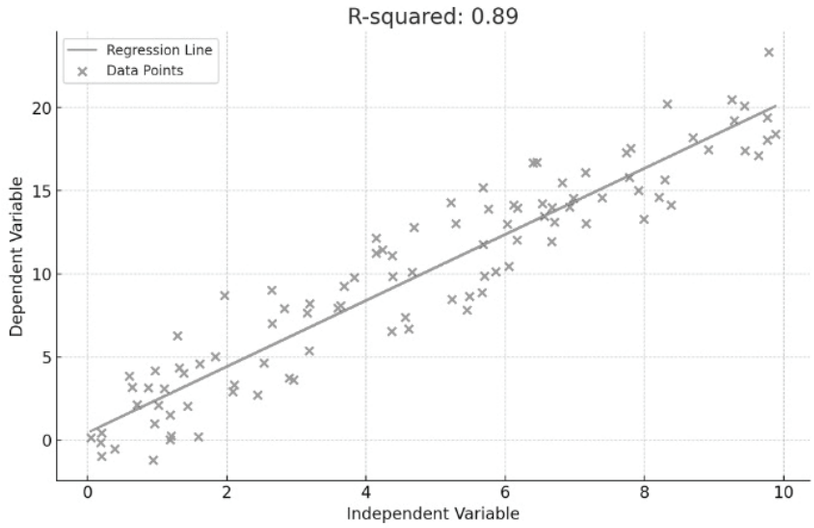


Figure 5.1: Graphical representation of R-square

5.5.2 *Interpretation of R-squared*

$0 \leq R^2 \leq 1$: R-squared values range from 0 to 1. A value of 0 indicates that the model does not explain any of the variability in the dependent variable, while a value of 1 indicates that the model explains all of the variability.

Notes

R^2 = 1: This implies a perfect fit where the model explains all the variability in the dependent variable.

R^2 = 0: This suggests that the model does not provide any improvement over using the mean of the dependent variable as a predictor.

```
Call:
lm(formula = y ~ ., data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.81177 -0.58567  0.05249  0.69674  2.40316

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04580    0.05694   0.804  0.42188
x1           0.42949    0.05874   7.311 2.52e-12 ***
x2           0.57386    0.06638   8.646 3.52e-16 ***
x3           0.26152    0.05773   4.530 8.58e-06 ***
x4          -0.29599    0.05444  -5.438 1.14e-07 ***
x5          -0.17564    0.05428  -3.236  0.00135 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9733 on 294 degrees of freedom
Multiple R-squared:  0.4131,    Adjusted R-squared:  0.4032
F-statistic: 41.39 on 5 and 294 DF,  p-value: < 2.2e-16
```

Figure 5.2: Results Showing R^2, *Adjusted* R^2 Values

It's important to note that while R-squared is a useful measure of goodness of fit, it has limitations. For instance, a high R-squared does not imply causation, and a low R-squared does not necessarily mean the model is useless. Additionally, R-squared may be sensitive to the inclusion or exclusion of variables in the model. Therefore, it is often recommended to consider other diagnostic measures along with R-squared when evaluating regression models.

5.6 Adjusted R Square/*Adjusted* R^2

The adjusted R-squared is a modified version of the R-squared statistic that adjusts for the number of predictors in a regression model. While R-squared measures the proportion of the variance in the dependent variable explained by the independent variables, adjusted R-squared penalizes the inclusion of irrelevant predictors that do not contribute significantly to the explanatory power of the model.

The formula for adjusted R-squared is:

$$Adjusted\ R^2 = 1 - \left(\frac{(1 - R^2)\,(n-1)}{n-k-1}\right)$$

Notes

where:

n is the number of observations.

k is the number of predictors (independent variables) in the model.

R^2 is the regular R-squared.

The adjustment in the formula is based on the number of predictors relative to the number of observations. The adjustment becomes larger as the number of predictors increases relative to the number of observations, penalizing models with a larger number of predictors.

Here are some key points about adjusted R-squared:

Penalty for Adding Predictors: The adjusted R-squared increases only if the new predictor improves the model more than would be expected by chance. If adding a new predictor does not improve the model significantly, the adjusted R-squared will decrease.

Maximum Value: Like the regular R-squared, the adjusted R-squared also ranges from 0 to 1. A higher adjusted R-squared suggests a better fit, but the penalty for including irrelevant predictors can prevent it from spuriously increasing.

Comparing Models: When comparing models with different numbers of predictors, the adjusted R-squared is often more informative than the regular R-squared. It helps to account for overfitting, which can occur when adding too many predictors that do not contribute meaningfully to the model's explanatory power.

In summary, the adjusted R-squared is a useful tool for assessing the goodness of fit of a regression model, especially when comparing models with different numbers of predictors. It provides a more realistic estimate of the model's explanatory power by penalizing the inclusion of unnecessary variables.

5.7 Standard Error of the Model

The Standard Error of the Model, also known as the standard error of the regression or standard error of the estimate, is a measure of the variability or dispersion of observed data points around the regression line in a regression analysis. It is a key component in assessing the precision of the regression model's predictions.

Notes

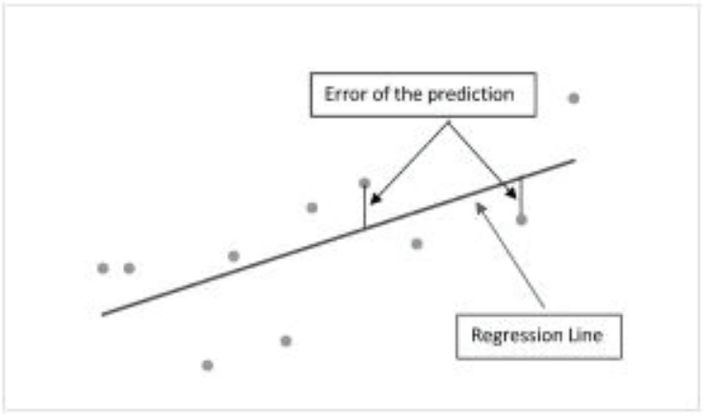


Figure 5.3: Graphical Representation of Standard Error

The formula for the Standard Error of the Model is:

$$Standard\ Error\ of\ the\ Model = \sqrt{\frac{Sum\ of\ square\ of\ residuals}{n-k}}$$

where:

Sum of Squares of Residuals (SSR): This represents the sum of the squared differences between the observed values and the values predicted by the regression model.

n: The number of observations.

k: The number of predictors (independent variables) in the model.

The standard error of the model gives an estimate of the standard deviation of the errors, or residuals, in the regression model. In other words, it quantifies the average amount by which the observed values deviate from the predicted values. A smaller standard error indicates that the model's predictions are relatively close to the observed values, suggesting a better fit.

Key points about the Standard Error of the Model are as follows:

Precision of Predictions: A lower standard error implies that the model's predictions are more precise and tend to be closer to the actual observed values.

Comparing Models: When comparing different regression models, the one with a lower standard error is generally preferred, as it indicates a better fit.

Notes

Degrees of Freedom Adjustment: The formula includes a degrees-of-freedom adjustment (dividing by $-k$) to account for the fact that estimating k parameters reduces the degrees of freedom.

Residual Standard Error: In some contexts, the term "residual standard error" is used interchangeably with the standard error of the model. They essentially represent the same concept.

In summary, the Standard Error of the Model provides a measure of the spread of residuals in a regression analysis. It helps assess the reliability and precision of the model's predictions and is often used in conjunction with other metrics for model evaluation.

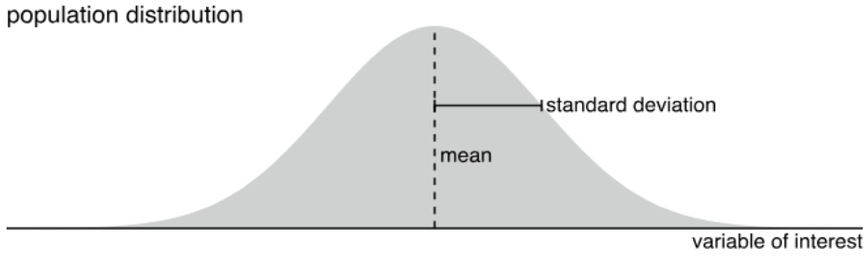


Figure 5.4: Distribution of model

5.8 Conceptual Understanding of AIC, BIC and SIC

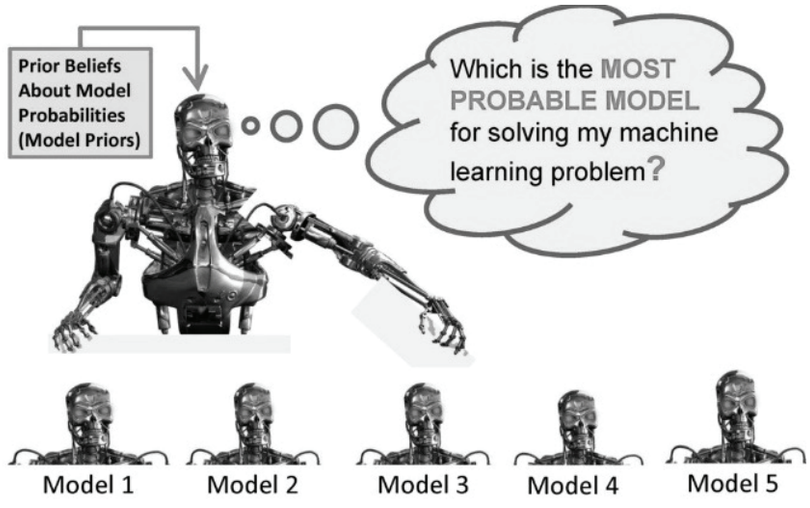


Figure 5.5: Finding an Answer to the Question "Which Model is Good Fit"

Notes

AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and SIC (Schwarz Information Criterion) are statistical measures used for model selection and goodness of fit. They are used to compare and choose among different models with the same dependent variable. These criteria help measure how well the models fit the given data. These criteria help balance the trade-off between the goodness of fit and the complexity of the model, penalizing overly complex models to avoid overfitting.

Akaike Information Criterion (AIC)

- Developed by Hirotugu Akaike, AIC is based on information theory.
- The formula for AIC is: AIC = –2 ln (L) + 2k, where L is the likelihood function and k is the number of parameters in the model.
- Lower AIC values indicate a better-fitting model. AIC penalizes models for having more parameters.

Bayesian Information Criterion (BIC)

- Also known as the Schwarz Information Criterion, BIC is a Bayesian-based criterion.
- The formula for BIC is: BIC = –2 ln (L) + k ln (n), where L is the likelihood function, k is the number of parameters and n is the sample size.
- Similar to AIC, lower BIC values indicate a better-fitting model. BIC penalizes complex models more strongly than AIC.

Schwarz Information Criterion (SIC)

- SIC is essentially the same as BIC and is often used interchangeably with BIC.
- Like BIC, SIC incorporates a penalty for the number of parameters in the model to prevent overfitting.

In summary, AIC, BIC, and SIC are all used to assess the trade-off between the goodness of fit and model complexity. Researchers and analysts typically compare these criteria for different models and choose the model with the lowest value, indicating the best compromise between fit and simplicity. It's important to note that while these criteria are useful, they are not without limitations, and their effectiveness may vary depending on the specific context and assumptions of the modelling process.

Notes

5.9 Calculation and Comparison of AIC, BIC and SIC

5.9.1 *AIC (Akaike Information Criterion)*

- **Principle:** AIC is based on information theory and aims to estimate the relative information lost when a given model is used to represent the true underlying process.
- **Penalization:** AIC penalizes models for having more parameters, but the penalty is relatively less severe compared to BIC.
- **Formula:** AIC = AIC = –2 ln (L) + 2k, where L is the likelihood function and k is the number of parameters in the model.
- **Selection Rule:** Choose the model with the lowest AIC value.

5.9.2 *BIC (Bayesian Information Criterion)*

- **Principle:** BIC is rooted in Bayesian probability theory and penalizes model complexity more strongly than AIC.
- **Penalization:** BIC penalizes models for the number of parameters and is more stringent than AIC, especially for smaller sample sizes.
- **Formula:** BIC = –2 ln (L) + k ln (n), where L is the likelihood function, k is the number of parameters and n is the sample size.
- **Selection Rule:** Choose the model with the lowest BIC value.

5.9.3 *SIC (Schwarz Information Criterion)*

- **Principle:** SIC is essentially the same as BIC and is often used interchangeably with it.
- **Penalization:** Like BIC, SIC penalizes models for the number of parameters and is designed to prevent overfitting.
- **Formula:** Similar to BIC.
- **Selection Rule:** Choose the model with the lowest SIC or BIC value.

Notes

5.9.4 *Model Selection Process*

All Possible Models

Ordered up to best 6 models up to 4 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC	Cp
sqft	1	0.6083	47.0060	1058.07	1065.64	44.8524
bath	1	0.4162	57.3859	1097.98	1105.54	113.9271
exempHS	1	0.2099	66.7577	1128.23	1135.80	188.0927
beds	1	0.0098	74.7336	1150.80	1158.37	260.0323
sqft,bath	2	0.6995	41.3806	1033.73	1043.73	14.0432
sqft,exempHS	2	0.6487	44.7453	1049.36	1059.36	32.3275
sqft,beds	2	0.6155	46.8106	1058.38	1068.38	44.2586
exempHS,bath	2	0.5205	52.2730	1080.46	1090.46	78.4080
bath,beds	2	0.4202	57.4808	1099.45	1109.45	114.4724
exempHS,beds	2	0.2142	66.9191	1129.86	1139.86	188.5552
sqft,exempHS,bath	3	0.7326	39.2395	1024.28	1036.67	4.1499
sqft,bath,beds	3	0.7043	41.2649	1034.35	1046.73	14.3323
sqft,exempHS,beds	3	0.6537	44.6576	1050.15	1062.54	32.5356
exempHS,bath,beds	3	0.5224	52.4423	1082.29	1094.67	79.7378
sqft,exempHS,bath,beds	4	0.7358	39.2089	1025.34	1040.07	5.0000

Figure 5.6: Mentioning the Model's AIC and BIC Values

All Possible Models

Ordered up to best 6 models up to 4 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC	Cp
sqft	1	0.6083	47.0060	1058.07	1065.64	44.8524
bath	1	0.4162	57.3859	1097.98	1105.54	113.9271
exempHS	1	0.2099	66.7577	1128.23	1135.80	188.0927
beds	1	0.0098	74.7336	1150.80	1158.37	260.0323
sqft,bath	2	0.6995	41.3806	1033.73	1043.73	14.0432
sqft,exempHS	2	0.6487	44.7453	1049.36	1059.36	32.3275
sqft,beds	2	0.6155	46.8106	1058.38	1068.38	44.2586
exempHS,bath	2	0.5205	52.2730	1080.46	1090.46	78.4080
bath,beds	2	0.4202	57.4808	1099.45	1109.45	114.4724
exempHS,beds	2	0.2142	66.9191	1129.86	1139.86	188.5552
sqft,exempHS,bath	3	0.7326	39.2395	1024.28	1036.67	4.1499
sqft,bath,beds	3	0.7043	41.2649	1034.35	1046.73	14.3323
sqft,exempHS,beds	3	0.6537	44.6576	1050.15	1062.54	32.5356
exempHS,bath,beds	3	0.5224	52.4423	1082.29	1094.67	79.7378
sqft,exempHS,bath,beds	4	0.7358	39.2089	1025.34	1040.07	5.0000

Figure 5.7: Choosing the Lowest AIC and BIC Values

- **Fit Multiple Models:** Start by fitting multiple models to your data, each representing a different hypothesis or complexity level.
- **Compute Information Criteria:** For each model, calculate the AIC, BIC, or SIC values using the respective formula.

Notes

- **Compare Criteria:** Compare the AIC, BIC, or SIC values for each model. Lower values indicate a better balance between goodness of fit and model complexity.
- **Select the Best Model:** Choose the model with the lowest AIC, BIC, or SIC value. This model is considered the most suitable given the trade-off between fit and complexity.
- **Consider Context:** While these criteria provide valuable guidance, it's important to consider the specific context of the analysis, the assumptions of the models, and the goals of the research.

In summary, the model selection process involves evaluating different models based on information criteria and choosing the one that strikes the best balance between goodness of fit and simplicity, as indicated by the lowest AIC, BIC, or SIC value.

IN-TEXT QUESTIONS

4. What is the primary purpose of AIC, BIC, and SIC in statistical model selection?
   (a) To measure the goodness of fit of a model.
   (b) To penalize complex models and avoid overfitting.
   (c) To determine the statistical power of a model.
   (d) To assess the multicollinearity among predictor variables.

5. Which of the following statements is true regarding AIC, BIC, and SIC?
   (a) Lower values indicate a better fit for all three criteria.
   (b) Higher values indicate a better fit for all three criteria.
   (c) AIC and BIC prefer simpler models, while SIC prefers more complex models.
   (d) AIC prefers more complex models, while BIC and SIC prefer simpler models.

6. Which of the following statements accurately describes the penalty term in BIC?
   (a) BIC penalizes the number of parameters linearly.
   (b) BIC penalizes the number of parameters logarithmically.
   (c) BIC penalizes the number of parameters quadratically.
   (d) BIC does not include a penalty term.

Notes

5.10 Summary

Goodness of fit is important for ensuring the validity and reliability of statistical models and analyses. A good fit means that the model or the distribution captures the essential features of the data and can be used for prediction or inference. The assessment of goodness of fit is crucial in determining whether a statistical model is appropriate for the given data.

If the goodness of fit is poor, it may indicate that the model needs improvement or that a different model should be considered. A poor fit means that the model or the distribution is not suitable for the data and may lead to erroneous or misleading conclusions.

On the other hand, a good fit suggests that the model is a reasonable representation of the underlying data patterns.

R-Square is a valuable metric in regression analysis, providing insights into how well the model explains the variability in the dependent variable. However, it should be used in conjunction with other evaluation metrics and considerations for a comprehensive assessment of the model's performance. The adjusted R-squared is a useful tool for assessing the goodness of fit of a regression model, especially when comparing models with different numbers of predictors. It provides a more realistic estimate of the model's explanatory power by penalizing the inclusion of unnecessary variables.

The Standard Error of the Model provides a measure of the spread of residuals in a regression analysis. It helps assess the reliability and precision of the model's predictions and is often used in conjunction with other metrics for model evaluation. AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and SIC (Schwarz Information Criterion) are statistical measures used for model selection and goodness of fit. They are used to compare and choose among different models with the same dependent variable. These criteria help measure how well the models fit the given data. These criteria help balance the trade-off between the goodness of fit and the complexity of the model, penalizing overly complex models to avoid overfitting.

Notes

5.11 Answers to In-Text Questions

1. *(a)* It measures how well the observed data matches the expected data in a statistical model

2. *(c)* Chi-square test

3. *(a)* Examining the difference between observed and expected values

4. *(b)* To penalize complex models and avoid overfitting

5. *(d)* AIC prefers more complex models, while BIC and SIC prefer simpler models

6. *(b)* BIC penalizes the number of parameters logarithmically

5.12 Self-Assessment Questions

1. Explain the significance of the Chi-Square test in assessing goodness of fit.
2. Explain the concept of residuals in context of regression modelling and goodness of fit.
3. What does the R-Square statistic measure in regression analysis, and how is it interpreted?
4. Define Goodness of Fit.
5. Why is overfitting a concern in the assessment of goodness of fit?

5.13 References

- David A. Freedman, Robert Pisani, & Roger Purves. (2009), "Statistical Models: Theory and Practice", Cambridge University Press, ISBN – 13: 978 - 0521743853.
- Ralph B. D'Agostino. (1986), "Goodness of Fit Techniques", CRC Press, ISBN: 13: 978-0824784052.
- Alan Agresti. (2018), "An Introduction to Categorical Data Analysis", Wiley, ISBN: 13: 978-1119405269.

Notes

5.14 Suggested Readings

- David S. Moore, George P. McCabe, & Bruce A. Craig. (2017). "Introduction to the Practice of Statistics", W. H. Freeman, ISBN-13: 978-1464158933.
- John Fox. (2015). "Applied Regression Analysis and Generalized Linear Models", Sage Publications, ISBN-13: 978-1452205663.

UNIT - IV

LESSON

6

Dummy Variables and Panel Data Regression Models

Dr. Tanu Kathuria
SDG Associate,
UNDP
Jammu & Kashmir, India
Email-Id: kathuriatanu@gmail.com

STRUCTURE

6.1 *Learning Objectives*
6.2 *Introduction*
6.3 *Concept of Dummy Variables*
6.4 *Types of Dummy Variables*
6.5 *Use of Dummy Variables to Model Qualitative/Binary/Structural Changes*
6.6 *Other Functional Forms of Dummy Variables*
6.7 *Response Regression Models*
6.8 *Panel Data Regression Model*
6.9 *Different Methods of Panel Data Estimation*
6.10 *Summary*
6.11 *Answers to In-Text Questions*
6.12 *Self-Assessment Questions*
6.13 *References*
6.14 *Suggested Readings*

Notes

6.1 Learning Objectives

- To determine how well the observed data fits with the model's expected value.
- To learn about different measures/tests/statistics used for goodness of fit.
- To understand the usefulness of R-Square and Adjusted R-Square.
- Learning about the Standard Error of the model with respect to Goodness of Fit.
- Conceptual understanding, comparison and calculation of AIC, BIC and SIC.

6.2 Introduction

Dummy variables, also known as indicator variables, are categorical variables used in statistical modeling and regression analysis to represent categorical data. These variables are binary, taking values of 0 or 1, and are often employed to capture qualitative information, such as group membership or the presence of a specific attribute.

In the context of regression analysis, dummy variables help incorporate categorical information into the model. For example, consider a categorical variable "Gender" with two categories, "Male" and "Female." To include this variable in a regression model, you can create a dummy variable that takes the value 1 if the observation is female and 0 if it is male. This allows the regression model to account for the impact of gender on the dependent variable.

The basic idea is to use dummy variables to represent different categories, and these variables become part of the regression equation as additional explanatory variables. When there are more than two categories, multiple dummy variables are created, with one less than the total number of categories to avoid multicollinearity.

Panel data, also known as longitudinal or time-series cross-sectional data, involves observing multiple entities over multiple time periods. Panel data regression models are designed to analyze such datasets, taking into account both the cross-sectional and time-series dimensions.

Notes

Panel data regression models offer several advantages, including the ability to control for individual heterogeneity, capture time trends, and handle endogeneity issues. These models are widely used in economics, finance, sociology, and other fields to analyze data with both temporal and cross-sectional dimensions.

6.3 Concept of Dummy Variables

Dummy variables, also known as indicator variables or binary variables, are used in statistical modeling and econometrics to represent categorical data numerically. They are called "dummy" because they take on the values of 0 or 1 to indicate the absence or presence of a particular categorical attribute. Dummy variables are especially useful when dealing with qualitative data that cannot be directly used in mathematical models. They are commonly used in regression analysis to represent categorial variables that have more than two levels, such as education level or occupation.

Here's how dummy variables work:

1. **Binary Representation:**
   - For a categorical variable with two categories (e.g., Yes/No, Male/Female), one dummy variable is sufficient. It takes the value of 0 for one category and 1 for the other.

Example: Gender

- Suppose you have a dataset with a "Gender" variable, and you want to include it in a regression model. You could create a dummy variable, say "Male," which takes the value 1 if the person is male and 0 if the person is female.

Gender	Male
Female	0
Male	1
Female	0
Male	1

2. **Handling Multiple Categories:**
   - For a categorical variable with more than two categories, you can create multiple dummy variables.

Notes

- If there are n categories, you typically create n−1 dummy variables. The excluded category becomes the reference category, and the other dummy variables represent whether an observation belongs to a specific category or not.

Example: Education Level

- Suppose you have an "Education Level" variable with categories like "High School," "Bachelor's," and "Master's." You can create two dummy variables, say "Bachelor's" and "Master's," and use "High School" as the reference category.

Education Level	Bachelor's	Master's
High School	0	0
Bachelor's	1	0
Master's	0	1
Bachelor's	1	0

3. **Avoiding Multicollinearity:**

- In regression analysis, including dummy variables for all categories without omitting one can lead to multicollinearity issues. Omitting one category helps prevent perfect correlation among the dummy variables.

Dummy variables are crucial for incorporating categorical information into statistical models, allowing analysts to account for the effects of different categories on the dependent variable in a meaningful way.

6.4 Types of Dummy Variables

Dummy variables, also known as indicator variables, are used to represent categorical data in regression analysis and statistical modeling. There are different types of dummy variables based on the number of categories in a categorical variable. Here are the main types:

1. **Binary Dummy Variables**

   This is the simplest type of dummy variable and is used for a categorical variable with two categories (binary). One category is chosen as the reference, and a single dummy variable is created to represent the other category. The dummy variable takes the value of 0 or 1, indicating the absence or presence of the category.

Notes

Example:

Let's say we have a variable "Gender" with categories Male and Female. A binary dummy variable, say D_{male}, would be created such that $D_{male} = 1$ for Male and $D_{male} = 0$ for Female.

2. Multicategory Dummy Variables

For categorical variables with more than two categories, multiple dummy variables are created. If a variable has k categories, $k - 1$ dummy variables are typically created, with one category chosen as the reference.

Example:

Consider a variable "Region" with categories North, South, and East. Two dummy variables, say D_{South} and D_{East} can be created. If both D_{South} and D_{East} are 0, it implies that the region is North (the reference category).

3. Interaction Dummy Variables

Interaction dummy variables are used when there are potential interactions between two or more categorical variables. These variables are created by taking the product of the dummy variables representing the individual categories.

Example:

If you have two categorical variables, A and B, with categories A1, A2 and B1, B2, interaction dummy variables like D_{A1B1}, D_{A1B2}, D_{A2B1}, and D_{A2B2} can be created to capture the joint effects.

These are the main types of dummy variables used in statistical modeling. The choice of which type to use depends on the nature of the categorical variable and the specific requirements of the analysis.

6.4.1 *Intercept Dummy Variables*

When using dummy variables to represent categorical variables, one category is usually chosen as the reference category, and dummy variables are created to represent the other categories. The intercept in a regression model represents the expected value of the dependent variable when all the predictor variables (including dummy variables) are set to zero.

Notes

Here's a brief explanation:

- **Reference Category:** One category of the categorical variable is chosen as the reference or baseline category. This category is not explicitly represented by a dummy variable.
- **Dummy Variables:** For a categorical variable with k categories, $k - 1$ dummy variables are created. Each dummy variable represents one of the non-reference categories. If the variable has k categories (including the reference category), then $k - 1$ dummy variables are created.
- **Intercept:** The intercept in the regression equation represents the expected value of the dependent variable when all predictor variables (including dummy variables) are set to zero. In the context of dummy variables, this means when the observation belongs to the reference category.

For example, suppose we have a categorical variable "Colour" with three categories: Red, Green, and Blue. If we create dummy variables for Green and Blue (with Red as the reference), the regression equation might look like this:

$$Dependent\ Variable = \beta_0 + \beta_1 * Green\ Dummy + \beta_2 * Blue\ Dummy + \varepsilon$$

In this equation:

β_0 represents the expected value when the colour is Red (the reference category).

β_1 represents the change in the expected value when the colour is Green.

β_2 represents the change in the expected value when the colour is Blue.

This setup allows for different intercepts for each category, and the coefficients associated with the dummy variables represent the differences from the reference category.

6.4.2 *Slope Dummy Variables*

A "slope dummy variable" typically refers to a dummy variable that is used to capture a difference in the slope of a regression line between different groups or categories. Dummy variables are binary (0 or 1) variables that are used to represent categorical data in regression models.

Notes

Here's a brief explanation:

- **Slope Dummy Variables:** A slope dummy variable would come into play when you suspect that the relationship between the independent variable (predictor) and the dependent variable (response) differs between the two groups. In other words, you want to allow for different slopes for the two groups.
- **Interaction Term:** The slope dummy variable is often used in interaction with the original independent variable. The interaction term is the product of the slope dummy variable and the original independent variable. This interaction term is added to the regression model to allow for different slopes for the two groups.

  The regression model with a slope dummy might look like this:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i * D_i) + \varepsilon_i$$

  In this equation:

  Y_i is the dependent variable.

  X_i is the original independent variable.

  D_i is the slope dummy variable (1 for Group B, 0 for Group A)

  ε_i is the error term.

β_0 is the intercept.

β_1 is the slope of Group A.

β_2 is the difference in intercepts between Group B and Group A.

β_3 is the difference in the slopes between Group B and Group A.

The presence of the interaction term ($X_i * D_i$) allows the slope of the regression line to vary between the two groups.

This approach is commonly used in Analysis of Variance (ANOVA) and regression analysis when there is a reason to believe that different groups may exhibit different relationships between variables.

6.4.3 *Interactive Dummy Variables*

Interactive dummy variables, also known as interaction terms or interaction dummy variables, are used to model how the effect of one variable on the dependent variable varies across different levels or groups

Notes

of another variable. These interaction terms allow for a more nuanced analysis of how the relationship between variables changes based on certain conditions.

For example, suppose you have two categorical variables: gender (male or female) and education level (high school or college). You are interested in understanding how the effect of education level on income differs between males and females. To capture this interaction, you would create an interaction term by multiplying the dummy variable for education level (let's call it $D_{college}$) by the dummy variable for gender (let's call it D_{female}). The interaction term would be $D_{college} * D_{female}$.

The regression model with interactive dummy might look like this:

$$Y_i = \beta_0 + \beta_1 D_{college} + \beta_2 D_{female} + \beta_3 (D_{college} * D_{female}) + \varepsilon_i$$

In this equation:

Y_i is the dependent variable.

$D_{college}$ is the dummy variable for college education (1 if the individual has a college education, 0 otherwise).

D_{female} is the dummy variable for gender (1 if the individual is female, 0 otherwise.

ε_i is the error term.

β_0 is the intercept.

β_1 is the effect of having a college education for males.

β_2 is the effect of being female for those with only a high school education.

β_3 is the additional effect on income for females with a college education compared to males with a college education.

The interactive dummy variable $D_{college} * D_{female}$ allows you to examine whether the effect of having a college education on income is different for females compared to males.

This approach is useful when you suspect that the relationship between two variables is not constant across different groups, and you want to account for this variation in your statistical model.

Notes

6.5 Use of Dummy Variables to Model Qualitative/ Binary/Structural Changes

Dummy variables are commonly used in statistical modeling, particularly in regression analysis, to represent qualitative or categorical data. They are particularly useful when dealing with binary or structural changes in the data. Here's how dummy variables are used for these purposes:

6.5.1 *Binary Variables*

Example: Consider a variable like "Gender," which has two categories: Male and Female.

Dummy Variable: Create a dummy variable (also known as an indicator variable) that takes the value 1 for one category (e.g., Female) and 0 for the other category (e.g., Male).

Regression Model: Include this dummy variable in your regression model. The coefficient associated with the dummy variable indicates the average change in the dependent variable when moving from one category to the other.

$$Dependent\ Variable = \beta_0 + \beta_1 * X_1 + \beta_2 * Gender\ Dummy + \varepsilon$$

Here, β_2 represents the average difference in the dependent variable between the two gender categories.

6.5.2 *Structural Changes*

Example: Imagine you have data for a period before and after the implementation of a new policy.

Dummy Variable: Create a dummy variable that takes the value 0 for the period before the policy change and 1 for the period after the policy change.

Regression Model: Include this dummy variable in your regression model to account for the structural change in the data.

$$Dependent\ Variable = \beta_0 + \beta_1 * X_1 + \beta_2 * Policy\ Dummy + \varepsilon$$

Notes

The coefficient β_2 now captures the average change in the dependent variable associated with the policy change.

6.5.3 *Interactions*

Sometimes, you may need to model interactions between dummy variables and other variables. For instance, the effect of a policy change might differ across different regions.

You can introduce interaction terms by multiplying two (or more) dummy variables. For example, an interaction between the policy change dummy and a regional dummy.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * Policy\ dummy + \beta_3 * Regional\ Dummy + \beta_4(Policy\ Dummy * Regional\ Dummy) + \varepsilon_i$$

The interaction term (β_4) captures how the effect of the policy change differs across regions.

Dummy variables are a powerful tool for handling categorical and structural changes in regression models, allowing for a more nuanced understanding of the relationships in the data.

6.6 Other Functional Forms of Dummy Variables

In addition to the basic use of dummy variables to represent binary categories or structural changes, there are several other functional forms and techniques involving dummy variables in statistical modeling. Here are a few examples:

A. Interaction Effects

Dummy variables can be used to model interaction effects between different categorical variables. For example, if you have two categorical variables A and B, you can create dummy variables for each category and an interaction term by multiplying the two dummy variables.

$$Y_i = \beta_0 + \beta_1 * Dummy\_A + \beta_2 * Dummy\_B + \beta_3(Dummy\_A * Dummy\_B) + \varepsilon_i$$

The interaction term (β_3) captures the joint effect of categories A and B.

Notes

B. Piecewise Linear Regression

Dummy variables can be used to model piecewise linear relationships. This is helpful when you expect different slopes or intercepts for different ranges of your independent variable.

$$Y_i = \beta_0 + \beta_1 * X_1 + \beta_2 * Dummy\_Indicator + \beta_3 * (X_1 * Dummy\_Indicator) + \varepsilon_i$$

Here, Dummy_Indicator takes the value 1 for observations within a specific range and 0 otherwise.

C. Polynomial Regression

Dummy variables can be used to model polynomial relationships. For instance, you might use dummy variables to represent different polynomial degrees.

$$Y_i = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \beta_3 * Dummy\_Indicator + \beta_4 * (X * Dummy\_Indicator) + \beta_5 * (X^2 * Dummy\_Indicator) + \varepsilon_i$$

Dummy_Indicator takes the value 1 for observations where the polynomial term is relevant and 0 otherwise.

D. Seasonal Dummy Variables

In time series analysis, dummy variables are often used to model seasonal effects. Each dummy variable represents a different season.

$$Y_i = \beta_0 + \beta_1 * X + \beta_2 * Dummy\_Winter + \beta_3 * Dumm\_Sping + \beta_4 Dummy\_Summer + \beta_5 * Dummy\_Fall + \varepsilon_i$$

The coefficients associated with seasonal dummy variables capture the average change in the dependent variable during each season.

These are just a few examples of how dummy variables can be used in various functional forms to capture different patterns and relationships in the data. The key is to think about the specific characteristics of your data and how dummy variables can best represent those patterns in your regression model.

6.7 Response Regression Models

Qualitative response regression models, also known as binary choice models or binary response models, are used when the dependent variable is categorical and takes on only two possible outcomes. The most common example is a binary outcome, such as "success" or "failure," "yes" or

Notes

"no," or "1" or "0". These models are widely employed in various fields, including economics, biology, medicine, and social sciences. Here are a few common qualitative response regression models:

6.7.1 *Logistic Regression*

Logistic Regression is a statistical method used for predicting the probability of a binary outcome. It's commonly used when the dependent variable is categorical and represents two classes, such as 0 or 1, yes or no, true or false. Logistic Regression models the probability that an instance belongs to a particular category.

The logistic function (sigmoid function) is at the core of logistic regression, ensuring that the predicted probabilities lie between 0 and 1. The logistic function is defined as:

$$P\,(Y = 1) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}}$$

Here:

$P\ (Y = 1)$ is the probability of the dependent variable being 1.

e is the base of the natural logarithm

β_0 is the intercept

$\beta_1, \beta_2 \ldots \beta_n$ are the coefficients associated with the independent variables $X_1, X_2 \ldots X_n$

The logistic regression model is estimated using a method called Maximum Likelihood Estimation (MLE). The goal is to find the values of the coefficients that maximize the likelihood of observing the given set of outcomes.

6.7.1.1 Key Characteristics and Considerations of Logistic Regression

Interpretation of Coefficients: The coefficients (β) represent the change in the log-odds of the dependent variable for a one-unit change in the corresponding independent variable.

Odds Ratio: The odds ratio is derived from the coefficients and represents the change in odds for a one-unit change in the independent variable.

Binary Outcome: Logistic regression is suitable for binary outcomes, but it can be extended to handle multinomial or ordinal outcomes in the form of multinomial logistic regression or ordered logistic regression.

Notes

Assumption of Linearity in Log-Odds: Logistic regression assumes a linear relationship between the independent variables and the log-odds of the dependent variable.

Diagnostic Measures: Model fit can be assessed using measures like the deviance, likelihood ratio tests, and the Hosmer-Lemeshow goodness-of-fit test.

Logistic regression is widely used in various fields, including medicine (predicting disease presence or absence), marketing (predicting customer churn), and social sciences (predicting voting behaviour). It's a powerful tool for binary classification problems when the relationship between the independent variables and the log-odds of the dependent variable is nonlinear.

6.7.2 *Probit Regression*

Probit regression is another statistical method used for modeling the probability of a binary outcome, like logistic regression. It is particularly common in econometrics and social sciences. Like logistic regression, probit regression is used when the dependent variable is binary and follows a probit link function.

The probit model assumes that the Cumulative Distribution Function (CDF) of a standard normal distribution is related to the probability of the binary outcome. The general form of the probit model is as follows:

$$P\,(Y = 1) = \emptyset\,(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$$

Here:

$P\ (Y = 1)$ is the probability of the dependent variable being 1.

ϕ is the cumulative distribution function of the standard normal distribution

β_0 is the intercept

$\beta_1, \beta_2 \ldots \beta_n$ are the coefficients associated with the independent variables $X_1, X_2 \ldots X_n$

In contrast to logistic regression, which uses the logistic function, probit regression uses the cumulative distribution function of the standard normal distribution. The probit model is estimated using methods such as Maximum Likelihood Estimation (MLE), similar to logistic regression.

Notes

6.7.2.1 *Key Characteristics and Considerations of Probit Regression*

Interpretation of Coefficients: Similar to logistic regression, the coefficients (β) in probit regression represent the change in the probability of the dependent variable being 1 for a one-unit change in the corresponding independent variable.

Odds Ratio: While logistic regression provides odds ratios, probit regression typically focuses on marginal effects, which represent the change in the probability of the dependent variable for a small change in the independent variable.

Link Function: The probit link function assumes a normal distribution for the latent variable, and the model is based on the standard normal distribution.

Nonlinear Relationship: Probit regression assumes a linear relationship between the independent variables and the latent variable, but the observed probabilities follow a nonlinear relationship.

Both probit regression and logistic regression are popular choices for binary outcome modeling. The choice between the two often depends on the specific context of the data and the underlying assumptions of the modeler.

6.8 Panel Data Regression Model

Panel data regression models are statistical models that are specifically designed to analyse data collected over time from multiple entities (cross-sectional units) such as individuals, firms, countries, etc. This type of dataset, known as panel data or longitudinal data, has both a time series and a cross-sectional dimension. Panel data regression models consider the individual variability and the temporal dynamics in the data.

6.8.1 *Importance of Panel Data*

Panel data, also known as longitudinal or panel dataset, is a type of dataset that observes multiple entities over multiple time periods. It consists of both cross-sectional and time-series dimensions, making it a valuable resource for researchers in various fields. The importance of panel data stems from several key advantages and applications:

Notes

Accounting for Heterogeneity: Panel data allows researchers to account for individual or entity-specific heterogeneity. By observing the same entities over time, one can control for unobserved characteristics that may vary across entities.

Dynamic Analysis: Panel data facilitates the analysis of dynamic processes and changes over time. Researchers can examine trends, patterns, and the evolution of variables within entities.

Efficiency and Statistical Power: Panel data often provides more efficient estimates compared to cross-sectional or time-series data alone. The increased sample size (observations across time and entities) enhances statistical power and allows for more precise parameter estimates.

Handling Endogeneity: Panel data allows researchers to address endogeneity issues more effectively. By including lagged values of variables or fixed effects, it is possible to control for potential feedback loops and endogenous relationships.

Reducing Selection Bias: Panel data can help mitigate selection bias. Observing the same entities over time reduces the risk of selecting entities with specific characteristics that may bias the analysis.

Testing for Causality: The longitudinal nature of panel data allows researchers to explore causal relationships more effectively. Techniques such as fixed effects or first-differencing can control for time-invariant unobserved factors.

Policy Evaluation: Panel data is essential for evaluating the impact of policy changes over time. Researchers can analyze how changes in policies affect outcomes within entities and compare this to a control group.

Modeling Heterogeneous Responses: Researchers can investigate how different entities respond to changes in independent variables. This is particularly useful when examining how economic, social, or environmental factors affect various entities differently.

Forecasting and Prediction: Panel data enables the development of forecasting models that take into account both temporal and cross-sectional dimensions. This can improve the accuracy of predictions compared to models based solely on cross-sectional or time-series data.

Cross-Sectional and Time-Series Analyses: Panel data allows for both cross-sectional and time-series analyses within the same dataset. Researchers

Notes

can study relationships across entities at a specific point in time or track changes within entities over time.

In summary, the importance of panel data lies in its ability to provide richer, more comprehensive insights into complex phenomena by combining cross-sectional and time-series information. This makes it a valuable tool for empirical research in economics, sociology, finance, epidemiology, and other disciplines.

6.8.2 *Assumptions of Panel Data*

When working with panel data, researchers typically make certain assumptions to justify the use of specific panel data regression models and ensure the validity of their analyses. Here are some common assumptions associated with panel data:

No Perfect Multicollinearity: Assumption: The independent variables in the model are not perfectly correlated with each other.

Rationale: Perfect multicollinearity makes it impossible to estimate unique coefficients for each variable.

Linearity: Assumption: The relationship between the dependent variable and the independent variables is linear.

Rationale: Panel data regression models assume a linear relationship for the parameters being estimated.

Independence of Errors: Assumption: The errors (residuals) are independent across entities and over time.

Rationale: Independence ensures that the observations in one entity or time period do not influence the errors of other observations.

Homoscedasticity: Assumption: The variance of the errors is constant across entities and over time.

Rationale: Homoscedasticity ensures that the spread of residuals is consistent, allowing for reliable inference.

No Serial Correlation: Assumption: The errors are not correlated across time periods for a given entity.

Notes

Rationale: Serial correlation indicates that the error in one period is related to the error in the preceding period, potentially biasing standard errors.

No Endogeneity: Assumption: The independent variables are not correlated with the error term.

Rationale: Endogeneity can lead to biased estimates, as the model assumes that the independent variables are exogenous.

Normality of Errors (for Some Models): Assumption: The errors follow a normal distribution.

Rationale: While some models, like the probit or logit models, assume normality of errors, many panel data models are relatively robust to deviations from normality due to the Central Limit Theorem.

Time-Invariant Individual Effects (For Fixed Effects Models): Assumption: The individual-specific effects are constant over time.

Rationale: Fixed effects models assume that the observed and unobserved factors affecting each entity do not vary over time.

No Measurement Error: Assumption: There is no measurement error in the independent or dependent variables.

Rationale: Measurement error can lead to biased estimates and affect the reliability of the results.

Stationarity (For Dynamic Models): Assumption: Variables involved in dynamic panel data models are stationary.

Rationale: Stationarity is essential for ensuring that the relationships estimated in the model are stable over time.

Random Sampling (If Applicable): Assumption: The panel data is a random sample from the population of interest.

Rationale: Random sampling ensures that the estimates are generalizable to the broader population.

Researchers should be aware of these assumptions and conduct diagnostic tests to assess their validity. Violations of these assumptions can lead to biased parameter estimates and affect the reliability of statistical inferences. Sensitivity analyses and robustness checks are important to evaluate the robustness of results to potential violations of these assumptions.

Notes

6.9 Different Methods of Panel Data Estimation

Here are some commonly used types of panel data regression models:

6.9.1 *Pooled OLS Regression*

The simplest approach is to pool all the data and use Ordinary Least Squares (OLS) regression. This treats the data as if it were a large cross-sectional dataset.

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + \beta_2 X_{it2} + \cdots + + \beta_k X_{itk} + \cup_{it}$$

where:

Y_{it} is the dependent variable for entity i at time t,

$X_{it1}, X_{it2}, \ldots, X_{itk}$ are the independent variables,

$\beta_0, \beta_1, \ldots \beta_k$ are coefficients, and

U_{it} is the error term.

6.9.2 *Fixed Effects Model*

Fixed effects models account for entity-specific effects that are constant over time. This model includes dummy variables for each entity to capture the unobserved heterogeneity.

$$Y_{it} = \alpha_i + \beta_1 X_{it1} + \beta_2 X_{it2} + \cdots + + \beta_k X_{itk} + \cup_{it}$$

where:

α_i represents the entity-specific fixed effect.

6.9.3 *Random Effects Model*

Random effects models assume that entity-specific effects are random and uncorrelated with the independent variables. It includes a random intercept term for each entity.

$$Y_{it} = \alpha + + \beta_1 X_{it1} + \beta_2 X_{it2} + \cdots + + \beta_k X_{itk} + \cup_{it}$$

Notes

where:

∞ is the overall intercept,

γ_i is the entity-specific random effect.

6.9.4 *First-Difference Model*

The first-difference model is often used to eliminate time-invariant individual effects. It involves differencing the data to remove individual-level heterogeneity.

$$\Delta Y_{it} = \beta_1 \Delta X_{it1} + \beta_2 \Delta X_{it2} + \cdots + + \beta_k \Delta X_{itk} + \cup_{it}$$

where:

Δ denotes the first difference.

6.9.5 *Dynamic Panel Data Models*

Dynamic models incorporate lagged values of the dependent variable and/or independent variables to account for time dependence and potential endogeneity.

$$Y_{it} = \rho Y_{it-1} + + \beta_1 X_{it1} + \beta_2 X_{it2} + \cdots + + \beta_k X_{itk} + \cup_{it}$$

where:

ρ captures the lagged effect.

Panel data regression models provide a flexible framework to analyze complex datasets by considering both the time and entity dimensions. The choice of model depends on the characteristics of the data and the specific research question being addressed.

IN-TEXT QUESTIONS

1. What is the purpose of using dummy variables in regression analysis?

(a) To increase the model complexity.

(b) To capture categorical information in the model.

(c) To reduce the number of observations.

(d) To introduce randomness.

Notes

2. How many dummy variables are typically created for a categorical variable with k categories in regression analysis?

(a) k − 1

(b) k

(c) k + 1

(d) 2k

3. In a regression model with a dummy variable representing gender (Male = 0, Female = 1), the coefficient associated with the dummy variable can be interpreted as:

(a) The average change in the dependent variable for males compared to females

(b) The probability of being a male

(c) The correlation between gender and the dependent variable

(d) The standard error of the dummy variable

4. What issue can arise if all dummy variables for a categorical variable are included in a regression model without omitting one?

(a) Perfect multicollinearity.

(b) Heteroscedasticity.

(c) Autocorrelation.

(d) Homoscedasticity.

5. Which panel data regression model includes entity-specific fixed effects?

(a) Pooled OLS regression

(b) Random effect model

(c) Dynamic panel data model

(d) Fixed effect model

6. In a fixed effect model for the panel data, how many dummy variables are typically created to represent entities?

(a) One

(b) Two

(c) n-1, where n is the number of entities

(d) n, where n is the number of entities

Notes

6.10 Summary

Dummy Variables: Dummy variables, also known as indicator variables, are binary variables that take on the values 0 or 1. They are used to represent categorical data or to incorporate qualitative information into regression models. Dummy variables help capture the effect of categorical variables that cannot be directly included in regression analysis.

For example, consider a categorical variable like gender (male or female). To include this in a regression model, a dummy variable can be created where 1 represents one category (e.g., male) and 0 represents the other category (e.g., female). This allows the regression model to account for the categorical nature of the variable.

Dummy variables are crucial in avoiding the issue of multicollinearity, where two or more independent variables are highly correlated. Including dummy variables for categorical factors helps prevent redundancy in the model.

Panel Data: Panel data, also known as longitudinal or cross-sectional time-series data, involves observations on a group of individuals, entities, or subjects over multiple time periods. This type of data structure allows for the analysis of both individual and time effects, providing richer information compared to purely cross-sectional or time-series data.

There are two main types of panel data:

1. **Balanced Panel:** This type of panel data includes the same set of entities observed over all time periods.
2. **Unbalanced Panel:** Here, not all entities are observed in every time period.

Panel data analysis enables researchers to account for individual-specific effects, time-specific effects, and interactions between individual and time effects. Common techniques include Fixed Effects Models and Random Effects Models, each with its own assumptions and implications.

In the context of panel data, dummy variables may be used to capture fixed effects, such as individual-specific characteristics that are constant over time.

In summary, both panel data and dummy variables are essential tools in econometrics and statistics. Panel data allows for the analysis of individual

Notes

and time effects, while dummy variables help incorporate categorical information into regression models, avoiding multicollinearity issues. Combining these concepts allows for a more comprehensive understanding of complex datasets.

6.11 Answers to In-Text Questions

1. *(b)* To capture categorical information in the model.

2. *(a)* k – 1

3. *(a)* The average change in the dependent variable for males compared to females.

4. *(a)* Perfect multicollinearity.

5. *(d)* Fixed Effect Model.

6. *(c)* n – 1, where n is the number of entities.

6.12 Self-Assessment Questions

1. Discuss the advantages of using panel data in empirical research. Provide examples of situations where panel data analysis is particularly beneficial.
2. What is the purpose of using panel data regression models in empirical research?
3. How can you interpret the coefficient of a dummy variable in a regression model?
4. What is the purpose of using dummy variables in regression analysis?
5. Consider a regression model with a categorical variable representing regions (North, South, East, West) using dummy variables. Describe how you would set up the dummy variables and interpret the results.

6.13 References

- Chamberlain, Gary. 1984. "Panel Data," in Griliches and Intriligator (eds.), Handbook of Econometrics, Volume 2. Amsterdam: North-Holland.

Notes

- Hsiao, Cheng. 1986. Analysis of Panel Data. Cambridge: Cambridge University Press.
- Gujarati, D. N. (2004). Basic Econometrics. 4th Edition, Mc-Graw Hill Companies, New York.

6.14 Suggested Readings

- Journal of Econometrics 18(1). 1982. Econometric Analysis of Longitudinal Data. (Especially the articles by Anderson and Hsiao, Chamberlain, and MaCurdy.)
- Mairesse, Jacques. 1990. "Time Series and Cross-Sectional Estimates on Panel Data: Why are They Different and Why Should They Be Equal?," in J. Hartog, G. Ridder, and J. Theeuwes (eds.), Panel Data and Labor Market Studies, North-Holland-Elsevier: 81–95.

Glossary

Adjusted R-Square: A modified version of R-Square that adjusts for the number of predictors in a regression model. It addresses the risk of overfitting by penalizing the inclusion of unnecessary variables.

Akaike Information Criterion (AIC): A measure that assesses the relative quality of statistical models. AIC balances the goodness of fit with the simplicity of the model, penalizing complex models to avoid overfitting.

Alternative Hypothesis: The hypothesis against which the null hypothesis is tested.

Assumptions: Conditions that must be met for the linear regression model to be valid, including linearity, independence of residuals, homoscedasticity, and normality of residuals.

Autocorrelation: Correlation between error terms in different time periods, violating the assumption of independence in panel data.

Autocorrelation: The correlation between the population disturbances across time period or space.

Auxiliary Regression: A regression used to compute a test statistic, used for the testing of heteroscedasticity or autocorrelation and not the estimation of the model of primary interest.

Balanced Panel: A panel dataset where the same set of entities is observed in each time period.

Bayesian Information Criterion (BIC): Similar to AIC, BIC is a criterion for model selection. It penalizes complex models more strongly than AIC, encouraging simplicity in the chosen model.

Best Linear Unbiased Estimator (BLUE): It is the estimator with the smallest variance, among all the linear unbiased estimators. Under the Gauss-Markov assumptions, the OLS estimators are BLUE.

Biased Estimator: It is an estimator whose expectation is not equal to the corresponding population parameter.

Binary Variable: A variable that can take on one of two possible values, often 0 or 1 in the context of dummy variables.

Chi-Square Test: A statistical test used to determine if there is a significant difference between the expected and observed frequencies in categorical data. Commonly employed in assessing goodness of fit for categorical models.

Notes

Classical Linear Model (CLM) Assumptions: These are the assumptions for multiple regression analysis. These assumptions include linearity of the parameters, no heteroscedasticity, no multicollinearity, zero conditional mean, no serial correlation and the normality of errors.

Coefficient of Determination (R-Square): A statistical measure representing the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. It ranges from 0 to 1, with higher values indicating a better fit.

Coefficients: Parameters in the linear regression equation representing the slope and intercept.

Confidence Interval: It is the random interval constructed so that in certain percentage of samples, the true parameter value lies in the interval. The percentage is decided according to the confidence level.

Confidence Level: It is the percentage of samples within which we want the confidence interval to contain the true parameter value.

Critical Value: In hypothesis testing, critical value refers to the value against which the estimated test statistics is compared to reject or not reject the null hypothesis.

Cross-Sectional Variation: Variation in the dependent variable across different entities at a single point in time

Degrees of Freedom: Under the multiple regression framework, it is the number of observations minus the number of estimated parameters.

Dependent Variable: The variable being predicted or explained in a regression model. It is denoted as Y in the context of linear regression.

Dependent Variable: The variable to be explained in a multiple regression model is the dependent variable.

Dummy Variable Trap: When all the categories of dummy variable are included in the regression along with the intercept term, then the model falls into a dummy variable trap.

Dummy Variable: A binary variable taking values 0 or 1 to represent different categories or groups in regression analysis.

Dummy Variable: It is a variable that takes values one or zero.

Durbin-Watson Statistic: It is a test statistic which is used to check for first-order serial autocorrelation in the classical linear regression model.

Notes

Econometric Model: It is an equation wherein dependent variable is related to a set of independent variables and unobserved disturbance term.

Econometrics: The application of statistical methods to economic data to test hypotheses and forecast future trends.

Error Term/Residual: Represents the variation in the dependent variable that is not explained by the independent variables in the model.

Error Term: In the multiple regression model, the error term is included to represent unobserved factors affecting the dependent variable.

Estimate: It is the numerical value of the estimator.

Estimator: It is a rule that estimates the population parameter using the sample data.

EVIEWS: Software specializing in time-series analysis and econometric modeling, commonly used for analyzing economic data.

Explanatory Variable: In multiple regression analysis, a variable that explains the variation in the dependent variable is called an explanatory or independent variable.

F-Distribution: It is a probability distribution obtained by forming the ratio of independent chi-square random variables, divided by their respective degrees of freedom.

Feature Engineering: The process of selecting, transforming, or creating new features to improve the performance of a regression model.

First Difference: It is the transformation of a variable obtained by taking the difference of values at adjacent time periods. The earlier period is subtracted from the later period.

Fixed Effects: Individual-specific effects in panel data analysis that are constant over time.

Forecast Error: The discrepancy between predicted and actual values, often used to assess the accuracy of forecasting models.

Forecasting: Predicting future values of economic variables based on historical data and model specifications.

Goodness of Fit: The degree to which a statistical model accurately represents the observed data. It measures how well the model fits the actual data points.

Notes

GRETL: Econometric software known for its user-friendly interface and capabilities in statistical analysis, modeling, and time-series analysis.

Heteroscedasticity: It is a condition when the variance of the error term is not constant, given the explanatory variable.

Heteroscedasticity: Unequal variance of the errors in a regression model, which can affect the efficiency of estimates.

Homoscedasticity: It is a condition when the variance of error term is constant, conditional on the independent variable.

Homoscedasticity: The assumption that the variance of residuals is constant across all levels of the independent variables.

Hypothesis Test: It is a statistical test of the null hypothesis, which is tested against the alternative hypothesis.

Hypothesis Testing: The statistical evaluation of hypotheses about the coefficients and the model's overall significance.

Independent Variable: The variable(s) used to predict the dependent variable. In multiple linear regression, there are multiple independent variables denoted as $X_1, X_2, ..., X_n$.

Indicator Variable: Another term for a dummy variable, used to indicate the presence or absence of a characteristic.

Interaction Term: A product of two or more variables, often used to capture joint effects in regression models.

Intercept (β_0): The constant term in the linear regression equation, representing the predicted value when all independent variables are zero.

Kolmogorov – Smirnov Test: A non-parametric test used to assess the goodness of fit of a sample to a distribution. It compares the sample distribution to a theoretical distribution.

Likelihood Ratio Test: A statistical test used to compare the fit of two nested models. It assesses whether the more complex model significantly improves the fit compared to the simpler model.

Linear Regression: A statistical modeling technique that establishes a linear relationship between a dependent variable and one or more independent variables.

Notes

Mean Squared Error (MSE): The average of the squared differences between observed and predicted values, used to evaluate the accuracy of a regression model.

Model Evaluation: Assessing the quality and validity of an econometric model by analyzing its predictive power and statistical significance.

Model Specification: The process of translating economic theories into mathematical equations representing relationships between variables.

Multicollinearity: A condition where two or more independent variables in a regression model are highly correlated, leading to issues in estimating the individual coefficients.

Multicollinearity: It is a situation when two or more explanatory variables in the regression model are *highly* correlated.

Multicollinearity: The presence of high correlation among independent variables in a regression model.

Multiple Linear Regression: A form of linear regression with more than one independent variable.

Null Hypothesis: In the hypothesis testing, null hypothesis is taken as true and require the data to provide statistical significance to reject it.

Ordinary Least Squares (OLS): It is a method of estimating the parameters in the multiple regression model, by means of minimizing the sum of squared residuals.

Outliers: Data points that significantly deviate from the overall pattern of the dataset and may influence the regression model disproportionately.

Outliers: These are the observations in the dataset which are different from most of the data points. These values are either very small or very large in comparison to the average value of the data.

Overall Significance of the Regression: It is obtained by testing the joint significance of all explanatory variables in the multiple regression model.

Panel Data: Data that includes observations on a group of entities over multiple time periods, allowing for the analysis of individual and time effects.

Parameter Estimation: The process of determining the numerical values of coefficients in an econometric model using statistical methods.

Notes

Perfect Collinearity: It is a situation in multiple regression model, wherein one explanatory variable is an exact function of the one or more explanatory variables.

Policy Analysis: Using econometric models to evaluate the potential impact of policy changes on economic variables and outcomes.

Pooling: A method of analyzing panel data by ignoring individual-specific effects and treating the data as a single cross-section.

Principal Component Analysis (PCA): A dimensionality reduction technique used to address multicollinearity by transforming correlated predictors into uncorrelated principal components.

R: An open-source programming language and software environment for statistical computing and graphics, extensively used in econometrics.

Random Effects: Unobserved individual-specific effects in panel data analysis assumed to be uncorrelated with the explanatory variables.

Reference Category: The category for which the dummy variable is set to 0. Other categories are compared to this reference category in regression analysis.

Regression Analysis: Statistical technique used to estimate the relationships between variables in a model, primarily through linear regression.

Regularization: Techniques like Ridge Regression and Lasso Regression introduce penalties on the size of regression coefficients to address over-fitting and multicollinearity.

Residual Analysis: Examination of residuals to assess the validity of linear regression assumptions, identify patterns, and detect outliers.

Residual sum of squares (RSS): It is the sum of the squared OLS residuals across all the observations in the data.

Residual: It is difference between the actual value of the dependent variable and the predicted value based on the regression model.

Residuals: The differences between observed and predicted values in a regression model.

Residuals: The differences between the observed values and the values predicted by a model. Residuals are used to assess the accuracy of the model's predictions.

Notes

R-squared (Coefficient of Determination): A metric indicating the proportion of variance in the dependent variable explained by the regression model.

Schwarz Information Criterion (SIC): Also known as the Schwarz Bayesian Criterion, it is an alternative to AIC and BIC for model selection. Like BIC, SIC penalizes complex models to a greater extent.

Simple Linear Regression: A form of linear regression with one independent variable.

Slope (β_1, β_2, …, β_n): The coefficients representing the change in the dependent variable for a one-unit change in the corresponding independent variable.

Standard Error of Regression (SER): In the multiple regression framework, the standard deviation of the population error is the standard error of regression. It is estimated as the square root of the sum of the squared residuals divided by the degrees of freedom.

STATA: Statistical software offering diverse capabilities for data analysis, modeling, and visualization, widely used in econometrics.

t-Distribution: This distribution is obtained as the ratio of a standard normal random variable and the square root of a independent chi-square random variable, wherein the denominator is divided by the degrees of freedom.

Time Effects: Systematic variations in the dependent variable over time, which may be captured in panel data analysis.

Time-Series Analysis: Methodology focusing on analyzing time-ordered data to understand patterns, trends, and relationships over time.

Unbalanced Panel: A panel dataset where not all entities are observed in every time period.

Unbiased Estimator: It is the estimator whose expected value is equal to the population value.

Variance Inflation Factor (VIF): A measure of how much the variance of an estimated regression coefficient increases if predictors are correlated.

Variance: It is the measure of the spread in the distribution of the random variable. It is calculated as the square of the standard deviation.

Department of Distance and Continuing Education
University of Delhi