Data Science for Economic and Social Issues

M. Jahangir Alam
Texas A&M University

January 30, 2024

Preface

This textbook is designed for students and researchers interested in the application of data science in economic and social issues...

Contents

List of Figures

List of Tables

Chapter 1

Introduction to Data Science and Economic Analysis

Contents

1.1 Introduction to Data Science Tools

1.2 Basics of Economic Policy Analysis

1.3 Quantitative Methods in Economics

1.4 Endogeneity and Selection Bias

1.4.1 Introduction to Endogeneity

Endogeneity is a significant concern in statistical modeling and econometrics. It refers to the scenario where key assumptions of the Classical Linear Regression Model (CLRM) are violated due to the correlation between the explanatory variables and the error term. The CLRM relies on several fundamental assumptions for the validity of the estimates:

- **Linearity**: The relationship between the dependent and independent variables is assumed to be linear.
- **Independence**: Observations are assumed to be independent of each other.
- **Homoscedasticity**: The error term is assumed to have a constant variance, irrespective of the value of the explanatory variables.
- **Normality**: For small sample sizes, it is assumed that the errors are normally distributed, at least approximately, for reliable inference.
- **No Endogeneity**: A critical assumption is that the error term should not be correlated with the independent variables.

Violation of these assumptions, particularly the absence of endogeneity, can lead to significant challenges in identifying causal relationships. Endogeneity can bias the estimates from a regression model, leading to incorrect conclusions about the relationship between the variables.

1.4.2 What is Endogeneity?

Endogeneity is a fundamental concept in econometrics that occurs when there is a correlation between an explanatory variable and the error term in a regression model. Consider a standard linear regression model:

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

In this model, Y represents the dependent variable, X is an explanatory variable, and u is the error term. Endogeneity is present if X is endogenous, which mathematically means that the covariance between X and u is not zero, i.e., $\text{Cov}(X, u) \neq 0$.

This correlation between the explanatory variable and the error term can arise from various sources such as omitted variables, measurement errors, or simultaneous causality. When endogeneity is present, it leads to biased and inconsistent estimators in regression analysis, posing significant challenges to drawing reliable conclusions about causal relationships.

1.5 Multicollinearity and Causality Identification

1.6 Approaches to Address These Problems

Chapter 2

Causal Estimation Techniques

Contents

2.1 Instrumental Variables (IV)

2.2 Difference-in-Differences (DID)

2.3 Regression Discontinuity Design (RDD)

2.4 Propensity Score Matching (PSM)

2.5 Interrupted Time Series (ITS)

Chapter 3

Data Handling and Machine Learning in Economics

Contents

3.1 Machine Learning Integration in Economics

3.2 Data Preprocessing and Visualization

3.3 Introduction to Prophet for Forecasting

3.4 Introduction to LSTM for Sequence Data Analysis

3.5 News Sentiment and Stock Price