Data Science in Economics and Social Sciences:

From Theory to AI Applications

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.

1.4.3 Types and Examples of Endogeneity

Endogeneity can manifest in various forms in econometric analyses, each with its unique implications. This subsection discusses the main types of endogeneity.

Omitted Variable Bias

Omitted Variable Bias occurs when a relevant variable that influences the dependent variable and is correlated with the independent variable is left out of the analysis. This can lead to a misestimation of the effect of the included independent variables. OVB arises because the omitted variable may be capturing some effects that are wrongly attributed to the included variables.

Simultaneity

Simultaneity arises when there is bidirectional causality between the dependent and independent variables. A classic example is the relationship between economic growth and investment. Economic growth can lead to increased investment (as profits and capital become more available), while higher investment can in turn boost economic growth. This two-way causation presents a simultaneity issue in the model.

Measurement Error

Measurement Error occurs when the variables in a model are measured with error. This leads to inaccuracies in estimating the relationship between the variables. When key variables are not measured accurately, it undermines the reliability of the model's estimations and can distort the actual impact of the variables.

1.4.4 Omitted Variable Bias

In econometric analyses, a common objective is to estimate the effect of certain variables on outcomes of interest. Consider a study designed to estimate the effect of class size (X) on student test scores (Y). A significant challenge in such analyses is the potential for omitted variable bias. This occurs when a variable that influences the dependent variable is not included in the model. For example, a student's family background might affect both the class size (X) and the test scores (Y), but it may not be included in the model.

The omission of such a variable can have critical implications. In this example, both the class size and the family background could independently affect the test scores. Neglecting to account for family background can lead to a misestimation of the true effect of class size on test scores. This bias occurs because the omitted variable (family background) captures part of the effect that is incorrectly attributed to class size.

This situation gives rise to endogeneity due to the correlation between the omitted variable (family background) and the included variable (class size). In the regression model, this correlation manifests as a correlation between the error term and the class size, leading to biased estimates. The implications of such a bias are far-reaching. The estimated effect of class size on test scores might be either overestimated or underestimated. Policy decisions based on these biased estimates could end up being ineffective or even counterproductive.

To mitigate omitted variable bias, several strategies can be employed. If data on the omitted variable (like family background in our example) is available, it should be included in the model. Alternatively, the use of instrumental variables that are correlated with the class size but not with the error term can help. Additionally, conducting a sensitivity analysis to assess the robustness of the results to the inclusion of potentially omitted variables can provide insights into the reliability of the findings.

In econometric analyses, understanding the structure of the regression equation is crucial, especially when dealing with omitted variable bias. Consider the following scenario:

The true model, which represents the actual relationship including all relevant variables, is given by:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon \tag{1.1}$$

In Equation (1.1), Y is the dependent variable, X and Z are independent variables, and ϵ is the error term. The inclusion of Z is essential to avoid bias in estimating the effect of X on Y.

However, the estimated model often omits crucial variables due to various limitations like data availability. This model might be represented as:

$$Y = \alpha_0 + \alpha_1 X + u \tag{1.2}$$

Omitting the variable Z in Equation (1.2) can lead to biased estimates of α_0 and α_1, particularly if Z is correlated with X and has an influence on Y.

To understand the impact of Z on the estimated model, consider expressing Z as a function of X:

$$Z = \gamma_0 + \gamma_1 X + \nu \tag{1.3}$$

This expression captures the part of Z that is and isn't explained by X.

When we substitute Equation (1.3) into the true model (Equation (1.1)), we obtain:

$$Y = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)X + (\epsilon + \beta_2\nu) \tag{1.4}$$

Equation (1.4) shows how the omitted variable Z affects the relationship between X and Y. The coefficient of X now reflects a combination of its direct effect on Y and the indirect effect via Z.

The error term in the estimated model (Equation (1.2)) becomes:

$$u = \epsilon + \beta_2\nu \tag{1.5}$$

This error term is compounded by the omitted variable Z's influence, captured by $\beta_2\nu$ in Equation (1.5). This leads to endogeneity, characterized by a non-zero covariance between X and u ($\text{Cov}(X, u) \neq 0$). Since u contains $\beta_2\nu$ and ν is associated with X (as Z is related to X), the error term becomes correlated with X. This correlation results in biased and inconsistent estimators.

The implication of this bias is significant, particularly in the interpretation of the effect of X on Y. The bias in the estimated coefficient α_1 in Equation (1.2) means it fails to provide an accurate estimate of the true effect β_1. This has serious implications in policy analysis and prediction, where accurate estimation of causal effects is critical.

The bias in the estimated coefficient of X, denoted as $\hat{\alpha}_1$, can be quantified as:

$$\text{Bias}(\hat{\alpha}_1) = \beta_2 \times \gamma_1 \tag{1.6}$$

Equation (1.6) shows that the omitted variable bias in the estimated coefficient of X is the product of the true effect of Z on Y (β_2) and the effect of X on Z (γ_1). This leads to a misrepresentation of the effect of X on Y, distorting the true understanding of the relationship between these variables. Such bias, if not addressed, can lead to misguided policy decisions. Accurate estimation, therefore, requires addressing this bias, potentially through the inclusion of Z in the model or via other statistical methods like instrumental variable analysis.

Mitigating omitted variable bias (OVB) is crucial for the accuracy and reliability of econometric analyses. There are several strategies to address this issue:

Including the omitted variable, when observable and available, directly addresses OVB by incorporating the previously omitted variable into the regression model. This approach is only feasible when the omitted variable is measurable and data are available. Including the variable not only reduces bias but also improves the model's explanatory power.

When the omitted variable cannot be directly measured or is unavailable, using proxy variables becomes an alternative strategy. A proxy variable, which is correlated with the omitted variable, can be used to represent it in the model. The proxy should ideally capture the core variation of the omitted variable. However, this method may not completely eliminate the bias, depending on how well the proxy represents the omitted variable.

The Instrumental Variables (IV) approach is another method used to mitigate OVB. In this approach, an instrumental variable is chosen that is uncorrelated with the error term but correlated with the endogenous explanatory variable (the variable affected by OVB). The IV approach helps in isolating the variation in the explanatory variable that is independent of the confounding effects caused by the omitted variable. The choice of a valid IV is crucial; it should influence the dependent variable only through its association with the endogenous explanatory variable. This method is commonly used in economics and social sciences, particularly when controlled experiments are not feasible.

Lastly, panel data and fixed effects models offer a solution when dealing with unobserved heterogeneity, where the omitted variable is constant over time but varies across entities, such as individuals or firms. These models help to control for time-invariant characteristics and isolate the effect of the variables of interest. Fixed effects models are especially useful for controlling for individual-specific traits that do not change over time and might be correlated with other explanatory variables.

Each of these methods has its strengths and limitations and must be carefully applied to ensure that the bias due to omitted variables is adequately addressed in econometric models.

1.4.5 Simultaneity in Economic Analysis

Understanding the relationship between economic growth and investment involves dealing with the issue of simultaneity. Economic growth can lead to more investment due to the availability of more profits and capital, while higher investment can, in turn, stimulate further economic growth. This mutual influence between economic growth and investment is a classic example of simultaneity, where each variable is endogenous, influencing and being influenced by the other. This simultaneous determination poses a significant challenge in identifying the causal direction and magnitude of impact between the two variables, making standard regression analysis inadequate.

Consider the investment function represented by:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1.7}$$

In Equation (1.7), Y represents investment, X economic growth, and ϵ the random error term capturing unobserved factors affecting investment.

The economic growth function can be modeled as:

$$X = \gamma_0 + \gamma_1 Y + \nu \tag{1.8}$$

Here, Equation (1.8) describes how economic growth X is influenced by investment Y, with ν as the random error term for unobserved factors affecting economic growth.

Substituting the economic growth function (Equation (1.8)) into the investment function (Equation (1.7)) gives us:

$$Y = \beta_0 + \beta_1(\gamma_0 + \gamma_1 Y + \nu) + \epsilon \tag{1.9}$$

Equation (1.9) illustrates the endogenous determination of Y and X, showcasing their interdependence.

Simplifying this equation, we obtain:

$$Y = \alpha_0 + \alpha_1 \nu + u \tag{1.10}$$

where $\alpha_0 = \beta_0 + \beta_1\gamma_0$, $\alpha_1 = 1 - \beta_1\gamma_1$, and $u = \epsilon + \beta_1\nu$. In this model, the error term u in Equation (1.10) is correlated with X since X is influenced by ν, and u includes ν. This correlation implies endogeneity and violates the Classical Linear Regression Model (CLRM) assumptions, leading to a biased estimator for β_1. The presence of endogeneity complicates the interpretation of regression results, particularly when assessing the effect of economic growth on investment.

The implications of this endogeneity are significant for econometric analysis. The bias in the estimator for β_1 means that conventional regression analysis will not accurately capture the true effect of economic growth on investment. This misrepresentation can lead to incorrect conclusions and potentially misguided policy decisions.

To address this issue, econometricians often resort to advanced techniques that can account for the simultaneity in the relationship between variables. One such approach is the use of structural models, where the simultaneous equations are estimated together, taking into account their interdependence. Another approach is using instrumental variables, where external variables that influence the endogenous explanatory variable but are not influenced by the error term in the equation are used to provide unbiased estimates.

The challenge, however, lies in correctly identifying and using these techniques, as they require strong assumptions and careful consideration of the underlying economic theory. Choosing an appropriate model or instrumental variable is crucial, as errors in these choices can lead to further biases and inaccuracies in the analysis.

In summary, simultaneity presents a complex challenge in econometric analysis, particularly in the study of relationships like that between economic growth and investment. Recognizing and addressing this simultaneity is key to uncovering the true nature of these economic relationships and providing reliable insights for policy-making and economic forecasting.

1.4.6 Measurement Error in Econometric Analysis

Understanding measurement error is crucial when estimating the effect of variables in econometric models. Consider a study aiming to estimate the effect of calorie intake (X) on weight gain (Y). Often, calorie intake is self-reported or estimated, leading to measurement errors. This error is typically not random and may be systematically biased due to underreporting or misreporting.

The nature of endogeneity in this context arises because the measured calorie intake (X^*) is $X^* = X + \text{error}$, where X is the true calorie intake, and 'error' represents the measurement error. The correlation between the true calorie intake (X) and the measurement error causes endogeneity. This correlation means that the error in X^* is related to X itself, violating the Ordinary Least Squares (OLS) assumption that the explanatory variables are uncorrelated with the error term.

The implications of this are significant. Estimates of the effect of calorie intake on weight gain using the mismeasured variable X^* will be biased and inconsistent. The direction of this bias depends on the nature of the measurement error, where systematic underreporting or overreporting can lead to an underestimation or overestimation of the true effect, respectively.

Mitigation strategies include using more accurate measurement methods for calorie intake, employing statistical techniques designed to address measurement error, such as Instrumental Variable (IV) methods, and conducting sensitivity analyses to understand the impact of potential measurement errors on the estimated effects.

The true model represents the actual relationship with the true, unobserved variable X^* and is given by:

$$Y = \beta_0 + \beta_1 X^* + \epsilon \tag{1.11}$$

Here, ϵ captures all other unobserved factors affecting Y. However, X is the observed variable, which includes the true variable X^* and a measurement error U:

$$X = X^* + U \tag{1.12}$$

Substituting the observed variable X into the true model gives us the substituted model:

$$Y = \beta_0 + \beta_1 X + (\epsilon - \beta_1 U) \tag{1.13}$$

The new error term now includes the measurement error. This leads to an altered error term in the regression with the observed variable:

$$\epsilon' = \epsilon - \beta_1 U \tag{1.14}$$

Since X includes U, and U is part of ϵ', ϵ' is correlated with X, violating the OLS assumption that the explanatory variable should be uncorrelated with the error term. This correlation leads to biased and inconsistent estimates of β_1, with the direction and magnitude of bias dependent on the nature of the measurement error and its relationship with the true variable.

The presence of measurement error, particularly in a key explanatory variable like calorie intake, can significantly distort the findings of a regression analysis. As illustrated in the substituted model (Equation (1.13)), the inclusion of the measurement error in the error term (Equation (1.14)) complicates the estimation process. The correlation of ϵ' with X, as per Equation (1.14), indicates that the standard Ordinary Least Squares (OLS) estimator will be biased and inconsistent, leading to unreliable estimates.

This bias in the estimator β_1 signifies that the estimated effect of calorie intake on weight gain, when relying on the mismeasured variable X, will not accurately reflect the true effect. The direction and magnitude of this bias are contingent upon the nature and extent of the measurement error. For instance, if the error is predominantly due to systematic underreporting, the estimated effect may be understated. Conversely, systematic overreporting could result in an overstated effect.

To mitigate the impact of measurement error, researchers must consider several strategies. Firstly, adopting more accurate methods to measure the key variables can significantly reduce the likelihood of measurement error. When direct measurement is challenging, using proxy variables that closely represent the true variable can be an alternative, though this approach may still retain some level of bias.

Moreover, the use of advanced econometric techniques, such as Instrumental Variable (IV) methods, provides a robust way to address endogeneity arising from measurement error. These methods rely on finding an instrument that is correlated with the mismeasured variable but uncorrelated with the error term, allowing for a more reliable estimation of the causal effect. However, finding a valid instrument can be challenging and requires careful consideration and validation.

Lastly, conducting sensitivity analyses is crucial to assess the robustness of the results to potential measurement errors. These analyses can help in understanding the extent to which measurement error might be influencing the estimated relationships and provide insights into the reliability of the conclusions drawn from the analysis.

In conclusion, measurement error poses a significant challenge in econometric modeling, particularly when key variables are prone to inaccuracies in measurement. Recognizing and addressing this issue is essential for ensuring the validity and reliability of econometric findings, especially in fields where accurate measurements are difficult to obtain.

1.4.7 Understanding Selection Bias

Selection bias is a critical issue in statistical analysis and econometrics, occurring when the samples in the data are not randomly selected. This bias arises in situations where the mechanism of data collection or the nature of the process being studied leads to a non-random subset of observations being analyzed. It is common in observational studies, especially in the social sciences and economics, where randomization is not always possible.

The presence of selection bias violates the assumptions of the Classical Linear Regression Model (CLRM), particularly the assumption that the error term is uncorrelated with the explanatory variables. This violation occurs because the non-random sample selection introduces a systematic relationship between the predictors and the error term, potentially leading to biased and misleading results in regression analysis. As a consequence, the estimates obtained may not accurately represent the true relationship in the population, which can lead to incorrect inferences and policy decisions.

Examples and common sources of selection bias include studies where participants self-select into a group or where data is only available for a specific subset of the population. For instance, in health studies examining the effect of diet on health, there may be an upward bias in the estimated diet effect if health-conscious individuals are more likely to participate. Similarly, in educational research, studying the impact of private schooling on achievement might lead to an overestimation of private schooling benefits if there is a selection of students based on parental dedication. Another example is in economic studies comparing earnings by education level, where college attendees are non-random and influenced by various factors, resulting in earnings comparisons being biased by unobserved factors like ability or background.

Mitigating selection bias involves employing techniques such as propensity score matching, instrumental variable analysis, or Heckman correction models. These methods aim to account for the non-random selection process and adjust the analysis accordingly. Additionally, ensuring a randomized selection process, if feasible, or accounting for the selection mechanism in the analysis, can help in reducing the impact of selection bias.

In summary, selection bias presents a significant challenge in statistical analysis, particularly in fields where controlled experiments are not feasible. Recognizing, understanding, and addressing this bias are essential steps in conducting robust and reliable econometric research.

1.4.8 Concluding Remarks: Navigating Endogeneity and Selection Bias

Navigating the complexities of endogeneity and selection bias is crucial in econometric models to ensure accurate causal inference and reliable research results. Endogeneity, a pervasive issue in econometrics, leads to biased and inconsistent estimators. It primarily arises from three sources: omitted variable bias, simultaneity, and measurement error. Each of these sources contributes to the distortion of the estimations in its way, making it imperative to understand

and address endogeneity comprehensively.

In parallel, selection bias presents a significant challenge in research, particularly when samples are non-randomly selected. This bias is common in various research contexts, ranging from health studies to education and economic research. It leads to misleading results that may not accurately reflect the true dynamics of the population or process under study. To mitigate the effects of selection bias, researchers must remain vigilant in their research design and employ appropriate techniques. Strategies such as propensity score matching and Heckman correction are commonly used to adjust for selection bias, especially in observational studies where randomization is not feasible.

Mitigation strategies for endogeneity include the use of instrumental variables, fixed effects models, and, where possible, randomized controlled trials. These methods aim to isolate the causal relationships and minimize the influence of confounding factors. Similarly, for addressing selection bias, techniques that account for the non-random selection process are essential. The overall significance of effectively dealing with these challenges cannot be overstated. Recognizing and appropriately addressing endogeneity and selection bias are fundamental to conducting robust and valid econometric analysis, leading to more accurate interpretations and sound policy implications.

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