Summer of Science (SoS) Mid-Term Report : Econometrics

Kavya Gupta

June 27, 2023

Contents

1 Week 1: Simple Regression Model

1.1 Introduction to the Simple Regression Model

The simple regression model forms the foundation of regression analysis. It allows us to examine the relationship between a dependent variable and a single independent variable. In the simple regression model, we assume a linear relationship between the two variables.

1.2 Assumptions of the Simple Regression Model

To apply the simple regression model, we make several assumptions:

- Linearity Assumption: The relationship between the dependent variable and the independent variable is linear.
- Independence Assumption: The errors or residuals are independent of each other.
- Homoscedasticity Assumption: The variance of the errors is constant across all levels of the independent variable.
- Normality Assumption: The errors follow a normal distribution.

1.3 Estimation Methods

The Ordinary Least Squares (OLS) method is commonly used to estimate the coefficients in the simple regression model. OLS minimizes the sum of squared differences between the observed values and the predicted values from the regression line.

1.4 Simple Regression Model Formula

The simple regression model can be represented by the following formulas:

$$PopulationModel: Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

$$SampleModel: Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \tag{2}$$

Where:

- Y is the dependent variable (population level).
- X is the independent variable (population level).
- β_0 is the intercept or constant term (population level).
- β_1 is the slope coefficient (population level).
- ε is the error term (population level).

- Y_i is the observed dependent variable in the sample.
- X_i is the observed independent variable in the sample.
- $\hat{\beta}_0$ is the estimated intercept or constant term.
- $\hat{\beta}_1$ is the estimated slope coefficient.
- e_i is the residual term or error in the sample.

1.5 Estimating Coefficients

The OLS estimation method is used to estimate the coefficients in the simple regression model:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{3}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \tag{4}$$

Where:

- n is the sample size.
- $\bar{X}$ is the mean of the independent variable in the sample.
- $\bar{Y}$ is the mean of the dependent variable in the sample.

1.6 Optimal Price for Maximal Turnover

In the context of pricing, we can utilize the simple regression model to determine the optimal price point that maximizes turnover or sales. By examining the relationship between price (independent variable) and turnover (dependent variable), we can estimate the slope coefficient $\hat{\beta}_1$ and identify the price that corresponds to the maximum turnover.

To find the optimal price for maximal turnover, we can set the derivative of the turnover equation with respect to price equal to zero:

$$\frac{d(Turnover)}{d(Price)} = \hat{\beta}_1 = 0 \tag{5}$$

Solving this equation will give us the optimal price for maximizing turnover.

2 Week 2: Multiple Regression

2.1 Introduction to Multiple Regression

In Week 2, we delve into the topic of Multiple Regression. Unlike Simple Regression, Multiple Regression allows us to examine the relationship between a dependent variable and multiple independent variables simultaneously. This allows for a more comprehensive analysis, considering the impact of multiple factors on the dependent variable.

2.2 Assumptions of Multiple Regression

The assumptions for Multiple Regression are similar to those of Simple Regression, with a few additional considerations:

- Linearity Assumption: The relationship between the dependent variable and the independent variables is linear.
- Independence Assumption: The errors or residuals are independent of each other.
- Homoscedasticity Assumption: The variance of the errors is constant across all levels of the independent variables.
- Normality Assumption: The errors follow a normal distribution.
- No Multicollinearity Assumption: The independent variables are not perfectly correlated with each other.

2.3 Estimation Methods

The estimation of coefficients in Multiple Regression is similar to that of Simple Regression. The Ordinary Least Squares (OLS) method is commonly used to estimate the coefficients, minimizing the sum of squared differences between the observed values and the predicted values from the regression equation.

2.4 Multiple Regression Model Formula

The Multiple Regression model can be represented by the following formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{6}$$

Alternatively, we can express the Multiple Regression model in matrix form as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{7}$$

Where:

- $\mathbf{Y}$ is the $n \times 1$ vector of dependent variables.

- $\mathbf{X}$ is the $n \times (k+1)$ matrix of independent variables, with an added column of ones for the intercept term.
- β is the $(k+1) \times 1$ vector of coefficients, including the intercept term.
- ε is the $n \times 1$ vector of errors or residuals.

2.5 Partial Effect

In Multiple Regression, the partial effect of an independent variable represents the change in the dependent variable associated with a one-unit change in that specific independent variable, holding all other independent variables constant. The partial effect can be calculated using the coefficient of the corresponding independent variable.

The partial effect formula for an independent variable X_i is given by:

$$\frac{\partial Y}{\partial X_i} = \beta_i \tag{8}$$

This represents the average change in the dependent variable for a one-unit change in X_i, holding all other independent variables constant.

2.6 Decomposition of Total Effect

The total effect of an independent variable on the dependent variable can be decomposed into two components: the direct effect and the indirect effect. The direct effect represents the immediate impact of the independent variable on the dependent variable, while the indirect effect captures the effect mediated through other independent variables.

The total effect can be calculated as the sum of the direct effect and indirect effect:

$$TotalEffect = DirectEffect + IndirectEffect \tag{9}$$

The direct effect represents the change in the dependent variable when the independent variable changes by one unit, while holding all other independent variables constant. It can be calculated using the coefficient of the corresponding independent variable.

The indirect effect represents the change in the dependent variable due to the indirect relationship mediated through other independent variables. It can be calculated by subtracting the direct effect from the total effect.

2.7 Interpreting Coefficients

The coefficients in Multiple Regression represent the average change in the dependent variable associated with a one-unit change in the corresponding independent variable, holding other independent variables constant.

The interpretation of coefficients depends on the nature of the independent variable. For example, if the independent variable is binary (0 or 1), the coefficient represents the difference in the dependent variable between the two groups.

2.8 Assessing Model Fit

To assess the overall fit of the Multiple Regression model, several statistical measures can be utilized, including the coefficient of determination (R^2), adjusted R^2, and the F-statistic. These measures help evaluate how well the model explains the variability in the dependent variable.

The coefficient of determination (R^2) represents the proportion of the variance in the dependent variable that can be explained by the independent variables. It is calculated as:

$$R^2 = \frac{SS_{explained}}{SS_{total}} \tag{10}$$

where $SS_{explained}$ is the explained sum of squares and SS_{total} is the total sum of squares.

The adjusted R^2 takes into account the number of independent variables and the sample size, providing a more accurate measure of the model's fit. It is calculated as:

$$Adjusted R^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k-1} \tag{11}$$

where n is the sample size and k is the number of independent variables.

The F-statistic is used to test the overall significance of the regression model. It compares the variation explained by the model to the unexplained variation. The F-statistic is calculated as:

$$F = \frac{MS_{explained}}{MS_{unexplained}} \tag{12}$$

where $MS_{explained}$ is the mean squared explained and $MS_{unexplained}$ is the mean squared unexplained.

2.9 Variable Selection

When dealing with multiple independent variables, it is important to consider variable selection techniques to identify the most influential variables. Techniques such as forward selection, backward elimination, and stepwise regression can aid in selecting the most relevant variables for the model.

These techniques involve iteratively adding or removing variables based on certain criteria, such as p-values, adjusted R^2, or information criteria like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).

3 Week 3: Model Specification

3.1 Introduction to Model Specification

In Week 3, we focus on the important topic of Model Specification. Model specification involves selecting the appropriate functional form and determining which variables to include in the regression model. A well-specified model ensures that the estimated coefficients are reliable and the model provides meaningful insights.

3.2 Functional Form

The functional form refers to the mathematical relationship between the dependent variable and the independent variables. It is crucial to choose an appropriate functional form that accurately represents the underlying economic theory or the data patterns.

For example, consider a simple linear regression model with a linear functional form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{13}$$

where Y is the dependent variable, $X_1, X_2, \ldots, X_k$ are the independent variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the coefficients, and ε is the error term.

3.3 Variable Transformation

Variable transformation involves modifying the original variables to better meet the assumptions of the regression model or to capture nonlinear relationships. Common variable transformations include taking logarithms, exponentiation, or adding polynomial terms.

For instance, consider a logarithmic transformation of a variable X:

$$X^* = \log(X) \tag{14}$$

This transformation can help address issues such as heteroscedasticity and nonlinearity in the relationship between X and the dependent variable.

3.4 Inclusion of Variables

Determining which variables to include in the regression model is an essential part of model specification. Considerations include theoretical relevance, statistical significance, and avoiding omitted variable bias.

It is important to include all relevant independent variables in the model to avoid omitted variable bias, which occurs when an important variable is left out of the regression model, leading to biased and inconsistent coefficient estimates.

3.5 Bias-Efficiency Tradeoff

In model specification, there is a tradeoff between bias and efficiency. Bias refers to the systematic deviation of the estimated coefficients from the true population coefficients. Efficiency, on the other hand, refers to the precision or reliability of the estimated coefficients.

When we include more variables in the regression model, the bias tends to decrease as the model becomes more flexible and can capture more complex relationships. However, increasing the number of variables can also lead to decreased efficiency, increasing the standard errors of the estimated coefficients.

3.6 Information Criteria

Information criteria, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), can be used to select the best-fitting model among competing specifications. These criteria balance the goodness of fit of the model with the complexity of the model.

The AIC is calculated as:

$$AIC = -2\log(L) + 2k \tag{15}$$

where L is the likelihood function of the model and k is the number of estimated parameters.

The BIC is calculated as:

$$BIC = -2\log(L) + k\log(n) \tag{16}$$

where n is the sample size.

Lower values of AIC and BIC indicate better-fitting models with a good tradeoff between fit and complexity.

3.7 Out-of-Sample Prediction

Out-of-sample prediction is the process of using the estimated regression model to predict the values of the dependent variable for observations not included in the original sample. This allows us to assess the predictive accuracy of the model.

3.8 Iterative Selection Methods

Iterative selection methods, such as forward selection, backward elimination, and stepwise selection, are used to systematically include or exclude variables from the regression model based on their statistical significance and contribution to the model fit.

Forward selection starts with an empty model and iteratively adds variables that improve the model fit the most. Backward elimination starts with a full model and iteratively removes variables that are least statistically significant.

Stepwise selection combines forward selection and backward elimination, allowing variables to enter or exit the model at each step based on predefined criteria.

3.9 RESET Test

The RESET (Regression Specification Error Test) is a diagnostic test used to detect functional form misspecification in the regression model. It tests whether adding higher-order terms of the predictors improves the model fit.

The null hypothesis of the RESET test is that the model is correctly specified, while the alternative hypothesis suggests functional form misspecification. The test calculates the F-statistic based on the residuals from the original model and their predicted values from a new model that includes additional terms.

3.10 Chow Break Test

The Chow Break Test is used to determine whether there is a structural change or break in the relationship between the independent variables and the dependent variable. It tests whether the coefficients of the variables differ significantly between two subsamples.

The null hypothesis of the Chow Break Test is that there is no structural change, while the alternative hypothesis suggests the presence of a break. The test calculates the F-statistic based on the residual sum of squares from the full model and the sum of squared residuals from separate models estimated on two subsamples.

3.11 Chow Forecast Test

The Chow Forecast Test is similar to the Chow Break Test but is specifically used to test the forecasting ability of a model. It examines whether separate models estimated on two subsamples provide significantly different forecasts.

The null hypothesis of the Chow Forecast Test is that there is no difference in forecasting ability between the two subsamples, while the alternative hypothesis suggests a difference. The test compares the mean squared forecast errors from separate models to determine if there is a significant difference.

4 Week 4: Endogeneity

4.1 Introduction to Endogeneity

In Week 4, we delve into the concept of Endogeneity, which arises when there is a correlation between the independent variables and the error term in a regression model. Endogeneity violates the assumption of exogeneity, leading to biased and inconsistent coefficient estimates.

4.2 Sources of Endogeneity

Endogeneity can occur due to various reasons, including omitted variable bias, simultaneity, and measurement error.

Omitted variable bias arises when a relevant variable is left out of the regression model, leading to a correlation between the omitted variable and the error term. This correlation biases the coefficient estimates of the included variables.

Simultaneity occurs when the dependent variable and one or more independent variables are jointly determined, creating a feedback loop. This violates the assumption of strict exogeneity and can lead to biased estimates.

Measurement error arises when the observed values of the variables are imprecise or subject to random errors. If the measurement error is correlated with the true values, it can introduce endogeneity in the regression model. The direction and magnitude of bias in the presence of measurement error depend on the type of measurement error: classical (non-stochastic) or Berkson (stochastic).

4.3 Instrumental Variables

Instrumental Variables (IV) estimation is a method used to address endogeneity in regression analysis. IV estimation relies on the use of instrumental variables, which are variables that are correlated with the endogenous variable but not directly with the error term.

The instrumental variables help establish a causal relationship between the independent variables and the dependent variable by isolating the variation in the independent variables that is not driven by the endogeneity issue.

4.4 Two-Stage Least Squares (2SLS)

Two-Stage Least Squares (2SLS) is a commonly used method for IV estimation. It involves two stages: the first stage estimates the relationship between the endogenous variable and the instrumental variables, and the second stage estimates the relationship between the dependent variable and the predicted values from the first stage.

The 2SLS estimator provides consistent and asymptotically efficient coefficient estimates under certain assumptions, such as relevance and exogeneity of the instruments. The estimation process involves formulating and solving a system of equations known as the formulation augmentation.

The matrix form of 2SLS estimation can be expressed as follows:

$$Y = X\beta + U \tag{17}$$

where Y is the dependent variable, X is the matrix of exogenous variables, β is the vector of coefficients, and U is the error term. The endogenous variable, denoted as Z, is replaced with its predicted values $\hat{Z}$ obtained from the first stage regression. The second stage regression is then performed as:

$$Y = X\beta + \hat{Z}\delta + \varepsilon \tag{18}$$

where δ represents the coefficients of the endogenous variables.

4.5 Asymptotic Properties

Under appropriate assumptions, the 2SLS estimator possesses desirable asymptotic properties. It is consistent, meaning that as the sample size increases, the estimator converges to the true parameter value. Additionally, it is asymptotically normally distributed, allowing for hypothesis testing and construction of confidence intervals.

The asymptotic distribution of the 2SLS estimator can be expressed as:

$$\sqrt{N}(\hat{\beta} - \beta) \to_d N(0, V) \tag{19}$$

where $\hat{\beta}$ is the estimated coefficient vector, β is the true coefficient vector, and V is the asymptotic variance-covariance matrix.

4.6 Solving Endogeneity: Graphical Representation

Graphical representation, such as scatterplots and correlation matrices, can help identify potential endogeneity issues. By visually inspecting the relationships between variables, we can detect patterns that suggest the presence of endogeneity. This graphical analysis can guide the selection of instrumental variables and the formulation of the 2SLS model.

4.7 Sargan Test

The Sargan test, also known as the overidentification test, is used to assess the validity of the instrumental variables in the 2SLS estimation. It tests whether the instrumental variables are uncorrelated with the error term in the model.

The Sargan test statistic is calculated as:

$$S = \hat{\varepsilon}' Z (Z'Z)^{-1} Z' \hat{\varepsilon} \tag{20}$$

where $\hat{\varepsilon}$ represents the residuals from the second stage regression and Z is the matrix of instrumental variables. The test statistic follows a chi-squared distribution with degrees of freedom equal to the number of overidentified restrictions.

4.8 Hausman Test

The Hausman test is another diagnostic test used to detect endogeneity in regression models. It compares the difference between the coefficients estimated by the 2SLS method (consistent but potentially inefficient) and the coefficients estimated by ordinary least squares (OLS) regression (potentially biased but efficient).

The Hausman test statistic is calculated as:

$$H = (\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'(V_{2SLS} - V_{OLS})^{-1}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}) \tag{21}$$

where $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{OLS}$ are the coefficient estimates from the 2SLS and OLS regressions, respectively, and V_{2SLS} and V_{OLS} are the corresponding variance-covariance matrices. The test statistic follows a chi-squared distribution with degrees of freedom equal to the number of endogenous variables.

4.9 Overestimation and Underestimation of OLS and Instruments

In the presence of endogeneity, ordinary least squares (OLS) estimation tends to overestimate the coefficients of the endogenous variables. This is due to the correlation between the endogenous variables and the error term.

On the other hand, instrumental variables (IV) estimation, such as 2SLS, provides consistent estimates of the coefficients by addressing endogeneity. However, if the instruments used are weak or irrelevant, IV estimation can lead to underestimation of the coefficients.

5 Conclusion

In Week 1, we covered the fundamentals of the simple regression model. We learned about its assumptions, estimation methods using OLS, and the formulas that represent the model. Additionally, we explored how the simple regression model can be applied to determine the optimal price for maximal turnover. This knowledge forms the basis for further exploration of more advanced regression techniques in the subsequent weeks of the course.

In Week 2, we explored the concept of Multiple Regression, which allows us to examine the relationship between a dependent variable and multiple independent variables. We discussed the assumptions, estimation methods, and the formulas that represent the Multiple Regression model, including its matrix form. Additionally, we examined the concepts of partial effect and the decomposition of total effect, interpreting coefficients, assessing model fit, and variable selection techniques. This knowledge provides a solid foundation for further exploration of advanced regression techniques in the subsequent weeks.

In Week 3, we explored the important concept of Model Specification. We discussed the selection of functional forms, the inclusion of variables, variable transformation, information criteria, out-of-sample prediction, and iterative selection methods. These concepts are essential for constructing reliable and meaningful regression models.

In Week 4, we explored the concept of Endogeneity and its implications for regression analysis. We discussed the sources of endogeneity, including omitted variable bias, simultaneity, and measurement error. We also introduced instrumental variables (IV) estimation, specifically the Two-Stage Least Squares (2SLS) method. Additionally, we discussed the formulation augmentation, asymptotic properties, solving endogeneity using graphical representation, and diagnostic tests such as the Sargan test and Hausman test. We also highlighted the overestimation and underestimation issues in OLS and instrumental variable estimation.
Endogeneity is a critical consideration in econometric analysis, and understanding how to detect and address endogeneity is vital for obtaining accurate and meaningful results.