# Summer of Science (SoS) Final Report : Econometrics

Kavya Gupta

July 26, 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:

$$\text{Population Model: } Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

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

Where:

- $Y$ is the dependent variable (population level).
- $X$ is the independent variable (population level).
- $\beta_0$ is the intercept or constant term (population level).
- $\beta_1$ is the slope coefficient (population level).
- $\varepsilon$ 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:

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

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:

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

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.
- $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector of coefficients, including the intercept term.
- $\boldsymbol{\varepsilon}$ 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:

$$\text{Total Effect} = \text{Direct Effect} + \text{Indirect Effect} \tag{9}$$

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:

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

where $SS_{\text{explained}}$ is the explained sum of squares and $SS_{\text{total}}$ 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:

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

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:

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

where $MS_{\text{explained}}$ is the mean squared explained and $MS_{\text{unexplained}}$ 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 $\varepsilon$ 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, $\beta$ 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 $\delta$ 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:

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

where $\hat{\beta}$ is the estimated coefficient vector, $\beta$ 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 Week 5: Binary Dependent Variables and Logit Models

## 5.1 Binary Dependent Variables

Binary dependent variables take on only two possible outcomes, typically coded as 0 and 1. Examples include yes/no, success/failure, and buy/sell. Binary data is common in many fields, and econometric models like the Logit model are used to analyze such data.

## 5.2 Linear Regression Model for Binary Dependent Variables

Using a linear regression model for binary dependent variables is not ideal as it can lead to predicted probabilities outside the [0, 1] range. Thus, we need a model that respects the nature of binary outcomes.

## 5.3 Logit Model

The Logit model is a type of binary response model that estimates the probability of an event occurring. It models the log-odds (logit) of the probability as a linear function of the independent variables.

The Logit model equation is given by:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \tag{22}$$

where:

- $p_i$ is the probability of the binary outcome for observation $i$.
- $\text{logit}(p_i)$ is the log-odds of the probability $p_i$.
- $\beta_0, \beta_1, \ldots, \beta_k$ are the coefficients of the independent variables $x_{i1}, x_{i2}, \ldots, x_{ik}$.

The logistic function (sigmoid function) is used to transform the logit back to the probability:

$$p_i = \frac{1}{1 + e^{-\text{logit}(p_i)}} \tag{23}$$

## 5.4 Odds Ratio

The odds ratio is an important concept in the Logit model. It measures the change in odds of an event occurring for a one-unit change in the independent variable.

The odds ratio is calculated as:

$$\text{Odds Ratio} = \exp(\beta_j) \tag{24}$$

where $\beta_j$ is the coefficient of the independent variable $x_j$.

## 5.5 Marginal Effect

The marginal effect in the Logit model measures the change in the probability of the binary outcome for a one-unit change in the independent variable.

The formula for the marginal effect of $x_j$ on $p_i$ is:

$$\text{Marginal Effect} = \frac{\partial p_i}{\partial x_j} = p_i(1 - p_i)\beta_j \tag{25}$$

where $\frac{\partial p_i}{\partial x_j}$ denotes the partial derivative of $p_i$ with respect to $x_j$.

## 5.6 Likelihood Function and its Construction

In the Logit model, we use the likelihood function to estimate the coefficients $\beta_j$ that maximize the likelihood of observing the binary outcomes given the independent variables.

The likelihood function for the Logit model is constructed as follows:

$$\mathcal{L}(\beta_0, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{N} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{26}$$

where $N$ is the number of observations, $y_i$ is the binary outcome (0 or 1) for observation $i$, and $p_i$ is the probability of the binary outcome as given by equation (1).

## 5.7 Maximum Likelihood Estimator and its Properties

The Maximum Likelihood Estimator (MLE) is used to estimate the coefficients in the Logit model that maximize the likelihood function.

The MLE properties include consistency, asymptotic normality, and efficiency. The MLE estimates have desirable statistical properties, such as being unbiased and having minimum variance among consistent estimators.

## 5.8 Covariance Matrix

The covariance matrix of the MLE estimates provides information about the precision and standard errors of the estimated coefficients. It is used to calculate confidence intervals and conduct hypothesis tests.

## 5.9 Logit Residuals

Logit residuals are used to assess the goodness-of-fit of the Logit model. They measure the difference between the observed and predicted probabilities.

## 5.10 Measures of Fit

Measures of fit evaluate how well the Logit model fits the observed data. Common measures include the likelihood ratio test, AIC (Akaike Information Criterion), and BIC (Bayesian Information Criterion).

## 5.11 Prediction Realization Table

The prediction realization table is a tool to assess the predictive performance of the Logit model. It compares the observed binary outcomes with the predicted outcomes from the Logit model and categorizes them into true positives, true negatives, false positives, and false negatives.

## 5.12 Overview

Week 5 focuses on Binary Dependent Variables and Logit Models, a crucial topic in econometrics. Binary dependent variables are outcomes with only two possible values, often represented as 0 and 1. Examples include whether a customer makes a purchase (1) or not (0) and whether a patient recovers from a disease (1) or not (0). Linear regression, commonly used for continuous outcomes, is unsuitable for binary data as it can produce predictions beyond the [0, 1] probability range. To address this, we explore the Logit model, which models the log-odds of the probability of the binary outcome as a linear function of the independent variables. We delve into the odds ratio, measuring the change in odds of an event occurring for a one-unit change in the independent variable, and the marginal effect, quantifying the change in the probability of the binary outcome. The likelihood function and Maximum Likelihood Estimator are introduced as key tools for estimating model coefficients. We also cover the covariance matrix, Logit residuals, measures of fit, and the prediction realization table for model evaluation. This week equips us with essential tools to analyze and make predictions about binary outcomes in real-world datasets, empowering us to make well-informed decisions in various domains.

# 6 Week 6: Time Series Analysis

## 6.1 Introduction to Time Series

Time series data is a sequence of observations collected over time. It is prevalent in various fields, such as finance, economics, and climate studies. Analyzing time series data allows us to identify patterns, trends, and seasonality, which are essential for forecasting future values.

## 6.2 Stationarity

Stationarity is a crucial concept in time series analysis. A time series is said to be stationary if its statistical properties, such as mean and variance, remain constant over time. Stationary time series exhibit no trend or seasonality, making it easier to model and forecast future values.

## 6.3 Autoregressive Model (AR)

The Autoregressive model is a popular time series model that predicts the value of a variable based on its past values. In the AR model, the current value is a linear combination of its previous values and a white noise error term.

The AR(p) model is given by:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + \varepsilon_t \tag{27}$$

where:

- $y_t$ is the value of the time series at time $t$.
- $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ are the past values of the time series.
- $\beta_0, \beta_1, \ldots, \beta_p$ are the coefficients to be estimated.
- $\varepsilon_t$ is the white noise error term.

## 6.4 Moving Average (MA)

The Moving Average model is another essential time series model that considers the relationship between the current value and the past forecast errors. In the MA model, the current value is a linear combination of the past forecast errors and a white noise error term.

The MA(q) model is given by:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \tag{28}$$

where:

- $y_t$ is the value of the time series at time $t$.
- $\mu$ is the mean of the time series.

- $\varepsilon_t, \varepsilon_{t-1}, \ldots, \varepsilon_{t-q}$ are the past forecast errors.
- $\theta_1, \theta_2, \ldots, \theta_q$ are the coefficients to be estimated.

## 6.5 Partial Autocorrelation Function (PACF)

The PACF is a tool to identify the order of autoregressive (AR) and moving average (MA) terms in a time series model. It measures the correlation between the current value and its lagged values, controlling for the intermediate lags.

## 6.6 Stochastic and Deterministic Trend

A time series can exhibit either a stochastic or deterministic trend. A stochastic trend indicates random fluctuations over time, while a deterministic trend shows a systematic increase or decrease over time.

## 6.7 Cointegration

Cointegration is a statistical property that allows two or more non-stationary time series to have a long-term relationship. Cointegrated series move together over time despite having individual trends.

## 6.8 Forecasting

Forecasting involves predicting future values of a time series based on its historical data and identified patterns. Various models, including ARIMA (AutoRegressive Integrated Moving Average) and Exponential Smoothing, are used for forecasting.

## 6.9 Granger Causality

Granger causality is a statistical concept used to test whether one time series can predict another. If the past values of one time series can improve the forecast of another, then there is Granger causality between them.

## 6.10 Consequences of Non-Stationarity

Non-stationary time series can lead to spurious relationships and inaccurate forecasts. Addressing non-stationarity is essential to obtain reliable results in time series analysis.

## 6.11 Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller (ADF) test is a statistical test used to determine whether a time series is stationary or not. It tests the null hypothesis of a unit root (non-stationarity) against the alternative of stationarity.

## 6.12 Error Correction Model (ECM)

The Error Correction Model (ECM) is a model that combines short-run dynamics of a time series with its long-run cointegrating relationship. ECM is often used when cointegrated series are involved.

## 6.13 Diagnostic Tests

Diagnostic tests help assess the adequacy of the chosen time series model. Common diagnostic tests include the Ljung-Box test for autocorrelation and the Jarque-Bera test for normality of residuals.

## 6.14 Breusch-Godfrey Test

The Breusch-Godfrey test is used to test for autocorrelation in the residuals of a time series model.

## 6.15 Engle-Granger Test for ECM

The Engle-Granger test is a procedure to test for cointegration between two non-stationary time series. It involves estimating a regression model and testing for the presence of a long-run relationship.

## 6.16 Overview

Week 6 delves into the captivating realm of Time Series Analysis, a powerful and versatile tool for understanding and predicting sequential data. We embark on a journey through the fundamental concepts and models in this domain. Starting with an introduction to Time Series, we explore the significance of stationarity, where statistical properties remain constant over time. Building on this, we examine the Autoregressive (AR) model, which predicts current values based on past values, and the Moving Average (MA) model, which considers past forecast errors. The Partial Autocorrelation Function (PACF) guides us in identifying the appropriate orders of these models. Additionally, we investigate the notion of stochastic and deterministic trends in time series data and explore cointegration, which allows for long-term relationships between non-stationary series. Forecasting, an indispensable aspect of time series analysis, is presented to predict future values with precision. We also delve into the concept of Granger causality, which determines whether one time series can predict another. Week 6 further addresses the consequences of non-stationarity and introduces the Augmented Dickey-Fuller (ADF) test for assessing stationarity. The Error Correction Model (ECM) is introduced to capture both short-run dynamics and long-run cointegrating relationships. Diagnostic tests, such as the Breusch-Godfrey test and the Engle-Granger test for ECM, are employed to validate chosen models. Armed with these techniques, we gain the expertise to analyze time series data, make informed forecasts, and unearth valuable insights in diverse domains, including economics, finance, and climate studies.

# 7 Week 7 : Econometrics and Indian GDP

In Weeks 7 and 8, we explore the practical applications of econometrics in the context of India. We investigate how econometric techniques and models play a vital role in understanding and analyzing the Indian economy, policy formulation, and decision-making processes.

Econometrics plays a crucial role in estimating and analyzing India's Gross Domestic Product (GDP). We delve into various methods used to estimate GDP, such as the production approach, expenditure approach, and income approach. Econometric models are employed to forecast GDP growth rates and understand the factors driving economic growth in India.

## 7.1 Production Approach

The production approach estimates GDP by summing up the value-added contributions of all sectors in the economy. We analyze sector-wise data and use econometric techniques to assess the contribution of agriculture, manufacturing, and services sectors to India's GDP.

## 7.2 Expenditure Approach

The expenditure approach calculates GDP by summing up the total expenditure on consumption, investment, government spending, and net exports. We utilize econometric models to study consumption patterns, investment trends, and external trade dynamics to estimate India's GDP using this approach.

## 7.3 Income Approach

The income approach estimates GDP by summing up all the incomes earned in the economy, such as wages, rents, profits, and interest. Econometric techniques are employed to analyze income data and derive GDP estimates.

## 7.4 GDP Calculation Example

Let's consider a hypothetical example of using the production approach to estimate India's GDP for the year 2023. We have the following data for the three sectors:

$$\begin{aligned} \text{Agriculture Value Added} &: \$100 \text{ billion} \\ \text{Manufacturing Value Added} &: \$150 \text{ billion} \\ \text{Services Value Added} &: \$200 \text{ billion} \end{aligned}$$

Using the production approach, we can calculate GDP as:

$$\text{GDP} = \text{Agriculture Value Added} + \text{Manufacturing Value Added} + \text{Services Value Added} \tag{29}$$

Substituting the values, we get:

$$\text{GDP} = \$100 \text{ billion} + \$150 \text{ billion} + \$200 \text{ billion} = \$450 \text{ billion} \tag{30}$$

Therefore, using the production approach, the estimated GDP for India in 2023 is \$450 billion.

## 7.5 GDP Growth Rate Forecast

Econometric models can also be used to forecast the GDP growth rate for India. Let's consider a simple ARIMA (AutoRegressive Integrated Moving Average) model to forecast the GDP growth rate for the next year (2024). Suppose we have historical GDP growth rate data for the last five years:

$$\begin{aligned}
&\text{Year 2019} : 5\% \\
&\text{Year 2020} : 4\% \\
&\text{Year 2021} : 6\% \\
&\text{Year 2022} : 5.5\% \\
&\text{Year 2023} : 6.5\%
\end{aligned}$$

We can use this data to estimate the parameters of the ARIMA model and forecast the GDP growth rate for 2024.

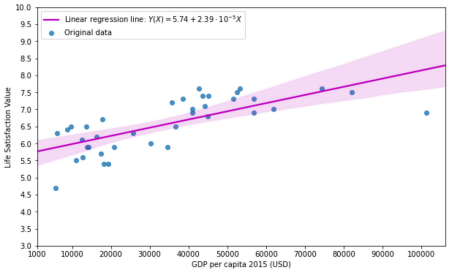


Figure 1: GDP Statistics

# 8 Week 8 : Econometrics and Minimum Wage Analysis

Econometrics is instrumental in analyzing the impact of minimum wage policies on the Indian labor market and the overall economy. We explore the effects of minimum wage changes on employment, wages, and income distribution.

## 8.1 Estimating Employment Effects

Using econometric models, we assess the effects of minimum wage changes on employment levels in various sectors of the Indian economy. We investigate whether minimum wage increases lead to job losses or job creation, and the potential implications for different industries.

## 8.2 Analyzing Wage Inflation

Econometrics helps us understand the impact of minimum wage adjustments on wage inflation in India. We study how minimum wage changes influence wage levels across different skill levels and regions, and their effect on the cost of living for low-income households.

## 8.3 Income Distribution and Poverty Analysis

We utilize econometric techniques to analyze the impact of minimum wage policies on income distribution and poverty levels in India. We investigate whether minimum wage adjustments lead to reduced income inequality and improved living standards for low-wage workers.

## 8.4 Minimum Wage Analysis Example

Let's consider a hypothetical example of analyzing the impact of a minimum wage increase in India. We have data on wage levels before and after the minimum wage hike:

Before Minimum Wage Hike:

$$\begin{aligned} \text{Average Wage} &: \$5 \text{ per hour} \\ \text{Number of Low-Wage Workers} &: 500,000 \end{aligned}$$

After Minimum Wage Hike:

$$\begin{aligned} \text{Average Wage} &: \$6 \text{ per hour} \\ \text{Number of Low-Wage Workers} &: 450,000 \end{aligned}$$

Using the data, we can calculate the total wages paid before and after the minimum wage hike:

Before Minimum Wage Hike:

$$\text{Total Wages Before} = \text{Average Wage} \times \text{Number of Low-Wage Workers} = \$5 \text{ per hour} \times 500,000 = \$2,500,000 \tag{31}$$

After Minimum Wage Hike:

$$\text{Total Wages After} = \text{Average Wage} \times \text{Number of Low-Wage Workers} = \$6 \text{ per hour} \times 450,000 = \$2,700,000 \tag{32}$$

By comparing the total wages before and after the minimum wage hike, we can analyze the impact on labor costs and the overall economy.

## 8.5 Graphs for Minimum Wage Analysis

In our minimum wage analysis, we can create several graphs to visualize the effects of minimum wage changes. For example, we can plot a line graph showing the trend in average wages before and after the minimum wage hike. Additionally, a bar graph can be used to compare the total wages paid to low-wage workers before and after the minimum wage increase. Another useful graph is a scatter plot, which can show the relationship between the minimum wage and employment levels in different industries.

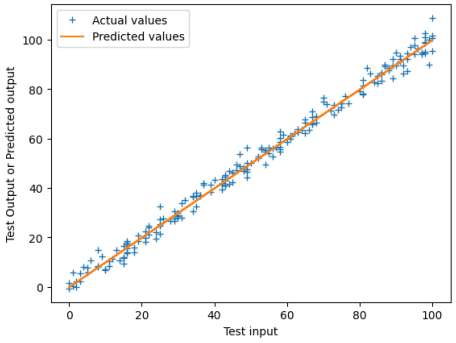


Figure 2: Graph1

## 8.6 Policy Implications and Decision Making

In Weeks 7 and 8, we also focus on the policy implications of econometric analyses in India. We assess how the results of econometric studies influence policy formulation related to economic growth, employment, inflation, and social welfare. Decision-makers in government, businesses, and organizations rely on econometric insights to make informed choices for the betterment of the Indian economy and its citizens.

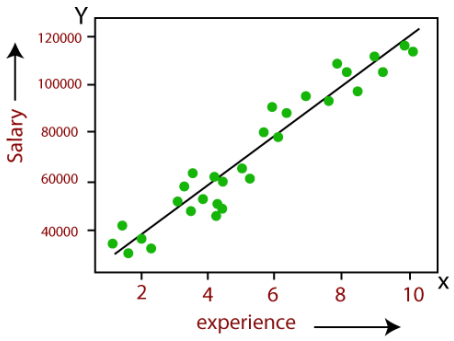


Figure 3: Graph2

# 9 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.

Week 5 has provided a comprehensive understanding of binary dependent variables and Logit models. The Logit model is a powerful tool for modeling and analyzing binary outcomes, such as yes/no or success/failure. We have learned how to estimate the model's coefficients using Maximum Likelihood Estimation and interpret the odds ratio and marginal effect for meaningful insights. Additionally, we have explored measures of fit and Logit residuals to assess the model's performance. Armed with this knowledge, we can apply Logit models to real-world datasets and make informed decisions in various fields, including finance, healthcare, and marketing.

Week 6 delves into the fascinating world of time series analysis. We have explored various time series models, including Autoregressive (AR) and Moving Average (MA) models, and the tools to identify their orders using the Partial Autocorrelation Function (PACF). Additionally, we have studied the concepts of cointegration and Granger causality, which are fundamental for understanding the relationships between multiple time series. To ensure accurate and reliable results, we have learned about diagnostic tests to validate the chosen time series models. Moreover, we have discussed the consequences of non-stationarity and how to address it using tests like the Augmented Dickey-Fuller (ADF) test. Armed with this knowledge, we can confidently analyze time series data, make meaningful forecasts, and discover valuable insights in various domains, including economics, finance, and climate studies

In Weeks 7 and 8, we delve into the fascinating realm of econometrics and its profound relevance in understanding and shaping the Indian economy. These weeks are dedicated to exploring the practical applications of econometric techniques, models, and analyses in the context of India's economic landscape.

Week 7 commences with a deep dive into the estimation and analysis of India's Gross Domestic Product (GDP). We explore various methods, such as the production approach, expenditure approach, and income approach, to estimate GDP. Econometric models come to the forefront in forecasting GDP growth rates and understanding the drivers behind India's economic growth.

Furthermore, we investigate the impact of minimum wage policies on the Indian labor market and the economy in Week 8. Econometrics plays a pivotal role in this analysis, as we assess the effects of minimum wage changes on employment levels, wage inflation, and income distribution. Through data-driven insights, we explore whether minimum wage adjustments lead to improved income equality and enhanced living standards for low-wage workers.

The weeks' discussions extend to the policy implications of econometric studies in India. Decision-makers in government, businesses, and organizations rely

on the valuable insights derived from econometric analyses to formulate effective economic policies, spur growth, and address socio-economic challenges.

In conclusion, Weeks 7 and 8 highlight the indispensable role of econometrics in understanding, analyzing, and shaping India's economic landscape. Armed with econometric tools and methodologies, we gain a deeper understanding of key economic indicators, make informed policy decisions, and work towards fostering sustainable economic growth and prosperity in our nation.