ELEMENTS OF ECONOMETRICS

B.A.(Economics) – Third Year

Paper Code : BAEC1934

PONDICHERRY UNIVERSITY

(A Central University)

DIRECTORATE OF DISTANCE EDUCATION

R.V. Nagar, Kalapet, Puducherry – 605 014

B.A III Year **Paper Code: BAEC 1934**

ELEMENTS OF ECONOMETRICS

Unit I: Nature and Scope of Econometrics

Meaning of econometrics – relationship between statistics, mathematics and economics – economic and econometric models – the aims and methodology of econometrics – historical origin of the term regression and its modern interpretation – statistical vs deterministic relationship – regression vs causation – regression vs correlation – terminology and notation – the nature and sources of data for econometric analysis.

Unit II: Two Variable and Multiple Regression Analysis

Unit III: The Problem of Inference

The normality assumption – hypothesis testing about individual partial regression coefficients – testing the overall significance of the sample regression – testing the equality of two regression coefficients

Unit IV: Assumptions of the Classical Regression Model

Assumptions of Classical Linear Regression Model- Introduction to the problems associated with the relaxation of the assumptions.

Unit V: Regression on Dummy Independent Variables

The nature of dummy variables – regression on one quantitative variable and one qualitative variable – regression on one quantitative variable and one qualitative variable with more than two classes – regression on one quantitative variable and two qualitative variables – interaction effects

References:

1. The Sustainability Revolution: Portrait of a Paradigm Shift by Edwards, Andres R., New Society Publishers, 2005.
2. Sustainable development in India: Stocktaking in the run up to Rio+20: Report prepared by TERI for MoEF, 2011.
3. Report of the Department for Policy Coordination and Sustainable Development (DPCSD), United Nations Division for Sustainable Development.
4. Corporate Social Responsibility Part I, Part II, Part III by David Crowther and Guler Aras
5. Weizsäcker, E. v. et al. (2009): Factor Five. Transforming the Global Economy Through 80 % Improvements in Resource Productivity. A Report to the Club of Rome, London, Sterling, VA (Earthscan)

TABLE OF CONTENTS			
UNIT	**LESSON**	**TITLE**	**PAGE NO.**
I	1.1	Meaning, Aims, and Methodology of Econometrics	1
	1.2	**Regression and Econometric Analysis**	13
II	2.1	**Simple Linear Regression Model**	41
	2.2	**Multiple Linear Regression**	72
III	3.1	**Evaluation of the Least Squares Estimates**	91
IV	4.1	**Stochastic Assumptions of the Classical Linear Regression Model**	125
	4.2	Nonstochastic Assumptions of the Classical Linear Regression Model	156
V	5.1	**The Nature of Dummy Variables**	195

UNIT – I : NATURE AND SCOPE OF ECONOMETRICS

Lesson 1.1 – Meaning, Aims, and Methodology of Econometrics

Structure

1.1.1 Meaning of Econometrics

Econometrics is the branch of economics that aims to give empirical content to economic relations. More precisely, it is the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.

Econometrics is the **science of using data to test and measure economic theories**. It combines **economics, statistics, and mathematics** to analyze real-world data and see if the theories hold up. **Economists** come up with ideas about how the economy works (theories). **Econometricians** then translate those ideas into **mathematical models** (like equations). They gather **real-world data** (like unemployment rates or stock prices). Using statistical tools, they run **experiments** with the data and the models. Based on the results, they can **confirm, refine, or even reject** the original theories.

Econometrics allows economists to test the predictions of economic theories against observed data. This is crucial for understanding whether a theory holds up in real life. Many economic theories predict relationships between variables (like income and consumption, or education and earnings). Econometrics provides tools to estimate the strength and form of these relationships. By understanding and quantifying economic relationships, econometrics can be used to make forecasts about future economic conditions or the outcomes of economic policies. Econometrics is often used to evaluate the effects of policy changes or to simulate the impact of potential policies before they are implemented.

1.1.2 Relationship Between Statistics, Mathematics, and Economics

Econometrics sits at the intersection of economics, mathematics, and statistics. It is essentially a tool that uses mathematical models to turn theories from economics into something that can be tested against real-world data using statistical methods. It draws from statistics and mathematics to provide a rigorous foundation for economic analysis. It is the conduit through which abstract economic theories are translated into practical tools for understanding, modeling, and predicting economic phenomena.

The relationship between statistics, mathematics, and economics is both intricate and symbiotic, especially when viewed through the lens of econometrics. Each discipline plays a crucial role in this collaborative endeavor, offering distinct yet intricately linked tools for empirical analysis.

At its heart, econometrics relies on **economic theories** to form hypotheses. These theories might be about anything from the way consumers make purchasing decisions to how entire economies respond to changes in policy. Econometrics uses **mathematical models** to express economic theories. These models help in structuring our understanding of complex relationships in a clear and concise way. Econometrics employs **statistical techniques** to test hypotheses and estimate the relationships described by economic theories. Through statistics, econometricians can determine if the relationships the theories predict are supported by real-world data.

Mathematics serves as the bridge between theoretical models and empirical analysis. Economic theories are often translated into mathematical equations, allowing for precise formulations of relationships between economic variables. Mathematics provides the language and framework necessary for formulating economic theories and models. It introduces precision and facilitates the expression of complex ideas in a form that can be systematically analyzed. In econometrics, mathematics is used to:

- **Model Economic Relationships:** Mathematical equations and functions are used to represent theories about economic relationships.
- **Optimization:** It is essential in solving problems related to maximization or minimization, such as cost minimization or profit maximization.
- **Structuring Problems:** Mathematics helps in structuring economic problems in a formal way, allowing for clear and concise representation of constraints, assumptions, and expected outcomes.

Statistics forms the bedrock of econometrics, providing the arsenal of methodologies for drawing meaningful inferences from data. It is crucial for testing hypotheses and estimating the relationships posited by economic theories. Without statistics, econometrics would lack the tools necessary to infer relationships from data. In econometrics, statistics is used to:

- **Data Analysis:** It helps in summarizing and extracting useful information from raw economic data.
- **Hypothesis Testing:** It provides methods for testing whether the observed data supports or refutes a given economic theory.
- **Estimation:** Statistical techniques such as regression analysis are used to estimate the parameters of economic models.

Economics provides the theories and concepts that econometrics aims to test and quantify. It is the substantive content that econometrics seeks to analyze using mathematics and statistics. In econometrics, economics provides:

- **Theoretical Framework:** The economic theories that are translated into mathematical models.
- **Context for Analysis:** It determines the variables that should be included in a model and the expected sign and size of relationships between variables.
- **Interpretation:** The economic theory helps in interpreting the results of statistical analysis, ensuring that they make sense within the broader economic context.

The interconnections between the three disciplines can be summarized as follows:

- **From Theory to Data:** Economics posits a theory, mathematics allows us to express this theory as a model, and statistics provides the tools to confront this model with data.
- **From Data to Theory:** Statistical results inform the refinement of economic theories, and mathematics helps to refine the models to better capture the complexities of economic behavior.

The triumvirate of statistics, mathematics, and economics, when combined in econometrics, allows for a systematic and empirical investigation of economic realities.

1.1.3 Economic and Econometric Models

In economics, models play a crucial role in understanding complex phenomena. They act as simplified representations of reality, allowing us to analyze and predict economic behavior. However, it is crucial to distinguish between two distinct types: economic models and econometric models. Economic models and econometric models are analytical frameworks that serve as tools to understand, interpret, and analyze economic phenomena and policy effects. Despite their interconnectedness, they serve distinct, yet complementary, roles within economic analysis.

Economic models are theoretical frameworks that describe the relationships between economic variables. They rely on assumptions about economic agents' behavior, market structures, and institutional arrangements. These assumptions abstract from the complexities of the real world, allowing economists to see how changes in one part of the system might have intended or unintended consequences in another. Common examples include supply and demand models, general equilibrium models, and macroeconomic models.

Mathematics is a fundamental tool in economic modeling, providing a language through which complex ideas can be expressed with clarity and precision. Economic models can be p**rescriptive**, suggesting how things should work under certain conditions (normative models), or **descriptive**, representing how things actually work in reality (positive models).

Econometric models, on the other hand, are statistical representations of economic relationships. They translate the theoretical framework of economic models into mathematical equations that can be estimated and tested using real-world data. This allows econometricians to assess the empirical validity of economic theories and quantify the relationships between variables. Examples include linear regression models, time series models, and panel data models.

Econometric models are statistical models that economists use to test hypotheses and forecast future economic activity. An econometric model is essentially the quantified version of an economic model, constructed with the intent of empirical testing. It involves **specification**, where the form of the econometric model is defined based on the theoretical model, **estimation**, where statistical methods are used to estimate the parameters of the econometric model, and **validation**, where the model is tested for its predictive power and its ability to explain the data. These models often

employ regression analysis, which is used to understand the relationship between a dependent variable and one or more independent variables, while accounting for randomness in the data.

The relationship between economic and econometric models is inherently cyclical and iterative. Economic models provide the theoretical basis for econometric models. They propose relationships that should be observed if the theory is correct. Econometric models then take these theoretical relationships and confront them with actual data to:

- **Test Economic Theories:** By estimating the parameters of the econometric model, researchers can test whether the relationships posited by economic theories hold in the real world.
- **Make Predictions:** Once validated, econometric models can be used to make predictions about the economy, which is particularly useful for policy analysis and forecasting.
- **Refine Theories:** The outcomes of econometric analyses can lead to the refinement of economic models. If real-world data do not support the theoretical model, economists may revise their models or question the underlying assumptions.

Here's an analogy: imagine an architect designing a building. The economic model is like the initial sketch, outlining the overall structure and function. The econometric model is like the detailed blueprint, incorporating specific measurements and materials based on real-world constraints. Just as the architect relies on both sketches and blueprints, economists rely on both economic and econometric models for a comprehensive understanding.

In essence, economic models hypothesize how the economy should work, while econometric models verify how the economy does work. Econometric models are the bridge between theory and data, allowing economists to go beyond abstract thought and to engage in empirical investigation, thereby validating and enhancing the understanding provided by economic models. By working together, they enable us to not only understand the "what" of economic phenomena but also quantify the "how much" and "how often," leading to deeper insights and informed decision-making.

1.1.4 Aims and Methodology of Econometrics

Econometrics, as a discipline, has a dual objective within the realm of economics: to provide quantitative content to theoretical relationships and to test economic theories against observed data. The methodology of econometrics is characterized by the application of statistical tools to economic data to lend empirical support to models and to forecast future trends.

Aims of Econometrics

Econometrics aims to test and refine economic theories so as to guide policy formulation and decision making. It has following objectives:

- **Estimation of Economic Relationships:** Econometrics seeks to estimate the parameters that define economic relationships. For instance, it aims to quantify the sensitivity of demand to changes in price or the effect of education on wages. These estimates allow economists to understand the magnitude and significance of relationships between economic variables.
- **Testing of Economic Theories:** A central aim of econometrics is to test hypotheses derived from economic theories. Econometricians employ empirical data to confirm or refute theoretical propositions, thereby providing a reality check for economic models.
- **Forecasting Economic Activity:** Econometrics is used to predict future economic outcomes based on the established relationships between variables. These forecasts are integral to decision-making by businesses, policymakers, and individuals.
- **Policy Analysis:** Econometric models are tools for policy analysis, allowing economists to simulate the effects of various policy options and to evaluate the impact of past policies. This aim ties closely with the concept of causal inference in econometrics, which is concerned with determining cause-and-effect relationships.

Econometrics empowers us to test and refine our understanding of economic phenomena, guiding policy decisions and shaping a more informed economic future.

Methodology of Econometrics

The methodology of econometrics encompasses several sequential steps:

- **Model Specification:** The process begins with the specification of an econometric model that is derived from economic theory. This involves choosing the appropriate variables and functional forms that are believed to best represent the economic reality being studied.
- **Data Collection:** Acquiring data that are relevant to the specified model is the next step. Econometricians may use cross-sectional, time-series, or panel data, depending on the nature of the economic phenomenon under investigation.
- **Estimation of Parameters:** Using statistical methods, econometricians estimate the parameters of the econometric model. Ordinary Least Squares (OLS) is the most common estimation technique for linear regression models, but there are many other methods, such as Maximum Likelihood Estimation (MLE), Generalized Method of Moments (GMM), and Instrumental Variables (IV) estimation, that are used when the OLS assumptions do not hold or when the model is not linear.
- **Hypothesis Testing:** After estimating the parameters, econometricians test various hypotheses related to the economic theory. This typically involves tests for statistical significance such as t-tests, F-tests, and chi-square tests.
- **Model Diagnostics:** Econometricians conduct diagnostic checks to assess the validity of the model. These checks may include tests for heteroskedasticity, autocorrelation, multicollinearity, and model misspecification.
- **Model Refinement:** Based on the outcomes of hypothesis testing and diagnostic checks, the econometric model may be refined to improve its accuracy and explanatory power.
- **Forecasting and Policy Simulation:** With a validated model, econometricians can make forecasts about future economic events or simulate the impact of potential economic policies.
- **Communication of Results:** The final step involves the interpretation and communication of the results in a manner that is understandable and useful for policymakers, economists, and other stakeholders.

The aims and methodology of econometrics are deeply intertwined. The discipline aims to validate and quantify economic theories using real-world data, which in turn informs economic decision-making and policy development. For example, suppose we are curious about the effect

of education on earnings. An economic theory might suggest that more education leads to higher earnings. To explore this, we would:

- **Develop a Model:** Create a mathematical representation of the theory, identifying the key variables (in this case, education and earnings) and how they are believed to interact.
- **Collect Data:** Gather data on individuals' education levels and their earnings.
- **Statistical Analysis:** Use statistical methods to analyze the data, testing whether the relationship between education and earnings is as the theory predicts, and if so, how strong that relationship is.
- **Interpret Results:** Assess the findings to see if they support the theory and consider what they might mean for policy or further research.

The methodology of econometrics, rooted in statistical analysis, is rigorous and iterative, ensuring that economic models are continually tested and refined against empirical evidence. By combining theory with data, econometrics provides insights into complex questions that matter for everything from individual choices to global economic policies.

1.1.5 Summary

Econometrics stands as the branch of economics that endows empirical significance to economic relations. It is defined by the quantitative dissection of genuine economic phenomena, an endeavor that marries theory with observation, all bound together through sound inferential methods. The discipline is tasked with estimating economic relationships, rigorously testing economic theories, forecasting economic behaviors, and evaluating as well as suggesting economic policies.

In the practice of econometrics, economic theories are not merely postulated but are transformed into mathematical models. These models are then populated with real-world data to assess whether the proposed theories withstand empirical scrutiny. The exercise involves using statistical instruments to probe and, where possible, predict economic tendencies and the consequences of policy measures. Through this process, theories can be either corroborated, fine-tuned, or contradicted by the collected data.

The interplay between statistics, mathematics, and economics in econometrics is intricate and mutually reinforcing. Mathematics serves as the scaffolding that allows economic theories to be structured into precise

mathematical models. These models are crucial for articulating complex economic relationships and solving problems related to optimization. Statistics underpins the endeavor, offering methodologies essential for extracting insights from data, testing hypotheses, and estimating the parameters that the mathematical models are based on. Economics, for its part, supplies the substantive theories and the contextual framework that guide the creation of econometric models and aid in the interpretation of statistical results.

The interconnection between these three fields is characterized by a dynamic flow from theory to data and back. Economic theories propose what relationships might exist, mathematics provides the means to express these relationships in a structured form, and statistics offers the tools to validate these theories against real data. When data speaks back to theory, it can lead to a refinement of the original economic propositions, ensuring that the models evolve to capture the complexities of economic behavior more accurately.

In distinguishing between economic and econometric models, one must appreciate that while the former are theoretical constructs that delineate the relationships between various economic variables based on certain assumptions, the latter are the statistical incarnations of these relationships. Econometric models translate theoretical frameworks into quantifiable equations that can be empirically tested with data. This allows for the assessment of the validity of economic theories and the quantification of economic interactions.

The methodology of econometrics is a disciplined process that starts with the specification of a model rooted in economic theory. It progresses with the collection of relevant data, followed by the statistical estimation of the model's parameters using a variety of techniques including Ordinary Least Squares and Maximum Likelihood Estimation. The model is then subjected to rigorous hypothesis testing, and its validity is gauged through diagnostic checks for potential issues such as heteroskedasticity and autocorrelation. Based on these evaluations, the model may undergo refinement to enhance its accuracy and explanatory power. With a model that withstands these tests, economists can forecast future economic events and simulate the impact of potential policies. The culmination of the process is the interpretation and communication of the results, which must be conveyed in a manner that informs policy decisions and furthers economic research.

1.1.6 Keywords

Causal Inference: The determination of cause-and-effect relationships within econometrics. This aspect is crucial for policy analysis, as it helps economists simulate the effects of various policy options and evaluate the impact of past policies using econometric models.

Econometrics: A branch of economics that applies statistical methods to economic data to empirically validate and quantify economic relationships. It integrates economics, mathematics, and statistics to test economic theories against real-world data and to estimate the strength and form of economic relationships.

Empirical Analysis: The process of using observed and collected data to test economic theories and hypotheses. In econometrics, this involves the use of statistical techniques to analyze data and evaluate the validity of economic models.

Empirical Grounding: The act of providing evidence based on real-world observations and data, as opposed to theoretical assumptions alone.

Heterogeneous Effects: The understanding that economic theories and policies can have different impacts on different groups or individuals within the economy.

Mathematical Models: Representations of economic theories in the form of mathematical equations. These models are used in econometrics to structure and formalize economic relationships, allowing for precise and systematic analysis.

Policy Simulation: The use of econometric models to predict the potential outcomes of different policy options before they are implemented.

Quantifiable Relationships: Economic connections that can be expressed as numerical values or formulas, allowing for precise measurement and analysis.

Regression Analysis: A statistical method used in econometrics to estimate the relationships between a dependent variable and one or more independent variables. It is commonly employed to test hypotheses and measure the impact of different factors on economic outcomes.

Statistical Validation: The process of using statistical methods to assess the accuracy and reliability of economic models, ensuring their conclusions are not due to chance.

1.1.7 Self-assessment Questions

1. What is the main objective of econometrics?
2. Explain the difference between economic models and econometric models.
3. How are statistics used in the field of econometrics?
4. What are some of the challenges involved in interpreting econometric results?
5. Can you think of an example where econometrics has been used to test an economic theory?
6. How could econometrics be used to evaluate the effectiveness of a government policy?
7. What are some of the limitations of using econometrics to make predictions about the future?
8. Imagine you are an economist studying the relationship between income and education. How might you use econometrics to analyze this relationship?
9. Why is it important to be critical of the assumptions made in econometric models?
10. Do you think econometrics can ever provide definitive answers to economic questions? Why or why not?

1.1.8 References

1. **Introductory Econometrics: A Modern Approach** by Jeffrey M. Wooldridge: This textbook provides a comprehensive introduction to econometrics, covering both theory and application. It's considered accessible for those with limited mathematical background.
2. **Econometrics** by Bruce D. Hansen: This textbook offers a more concise and theoretical treatment of econometrics, suitable for undergraduates with a strong grasp of statistics and mathematics.
3. **Using Econometrics: A Gentle Introduction** by A. Colin Cameron and Pravin K. Trivedi: This textbook focuses on the practical application of econometrics, presenting various econometric techniques through real-world examples.

4. **Mostly Harmless Econometrics: An Introduction in Plain English** by Joshua Angrist and Jörn-Steffen Pischke: This book uses humor and real-world examples to explain complex econometric concepts, making it a good choice for those seeking a more engaging learning experience.
5. **The Econometrics of Macroeconomic Policy** by Jordi Galí: This book delves into the specific application of econometrics in macroeconomic analysis, requiring a strong understanding of both statistics and economic theory.

Lesson 1.2 – Regression and Econometric Analysis

Structure

1.2.1 Historical Origin of the Term Regression

Regression, in statistical terms, is a methodological framework employed to discern the relationship between a dependent variable and one or more independent variables. A cornerstone of statistical analyses, it boasts a rich history interwoven with biological observations, astronomical calculations, and the search for meaning in diverse datasets. The term "regression" first emerged in the 19th century through the work of Sir Francis Galton, a polymath fascinated by heredity and biometrics. While studying pea plant heights, he noticed a curious phenomenon: the heights of offspring from tall parents tended to be closer to the average population height than their parents, exhibiting a "regression" towards the mean. He termed this phenomenon "regression towards mediocrity," later shortened to "regression." This observation led him to define "regression" as the tendency of offspring to be less extreme than their parents.

While Galton coined the term, the underlying mathematics had already established roots. In 1805, Adrien-Marie Legendre and Carl Friedrich Gauss, independently, published the method of least squares. This method sought to find a line that best fit a set of data points by minimizing the sum of the squared distances between the data points and the line. This

technique, initially used for analyzing planetary movements, proved fundamental for regression analysis. Adolphe Quetelet (1846) popularized and applied least squares extensively in social sciences, paving the way for broader statistical applications. The 20th century saw substantial advancements in formalizing regression theory. Karl Pearson (1894) built upon Galton's work, introducing the correlation coefficient to quantify the strength of linear relationships. This coefficient measures the strength and direction of the linear relationship between two variables, enabling more nuanced analyses beyond simple regression to the mean. This paved the way for multiple regression, allowing analysts to consider the influence of multiple independent variables on a dependent variable. Ronald Aylmer Fisher (1925) laid the groundwork for modern regression analysis with his seminal work on analysis of variance (ANOVA) and maximum likelihood estimation.

Regression analysis found applications in diverse fields, including econometrics--estimating relationships between economic variables like income, prices, and demand, social sciences--analyzing relationships between social factors like education, income, and crime rates, and natural sciences--modeling relationships between physical phenomena like temperature, pressure, and chemical reactions. Econometrics, in particular, embraced regression with enthusiasm. Ragnar Frisch (1934) developed the Cowles Commission methodology, establishing regression as a core tool for economic modeling and forecasting. Trygve Haavelmo (1944) introduced the concept of simultaneous equations, addressing the issue of interdependent variables in economic systems. Furthermore, econometrics extensively utilizes regression analysis to build econometric models that simulate and forecast economic behavior. These models play a crucial role in policymaking and understanding economic trends.

The post-war era witnessed a surge in computing power, making regression accessible to a wider audience. With the advent of powerful computers and advanced statistical software, regression analysis has become even more sophisticated. Software packages like SAS and SPSS facilitated ease of use, propelling regression into diverse fields, from finance and marketing to medicine and social sciences. Advanced techniques like logistic regression and non-linear regression emerged, providing greater flexibility and handling complex data structures and relationships.

The historical evolution of regression has also involved the development of robust statistical techniques. These techniques are designed to be

less sensitive to outliers and other forms of data that do not meet the assumptions of classical regression models. The use of bootstrapping and other resampling techniques has also improved the estimation of regression models by allowing for the estimation of the sampling distribution of almost any statistic.

The expansion of regression analysis over time has been characterized by both broadening the types of relationships it can model and deepening the theoretical underpinnings of these models. This has entailed not only advancements in the types of regression models available but also improvements in the underlying estimation techniques and diagnostic measures.

From the mid-20th century onward, the development of regression models has been influenced by the rise of computational power and the advent of new statistical theories. The proliferation of computers and statistical software in the latter half of the 20th century precipitated the development of numerous types of regression techniques beyond the basic linear model. The introduction of Generalized Linear Models (GLMs) by Nelder and Wedderburn in 1972 was a significant milestone. GLMs extend the linear model framework to allow for response variables that have error distribution models other than a normal distribution. These include, but are not limited to, logistic regression for binary outcomes, Poisson regression for count data, and Cox regression for survival analysis. These techniques are widely used in fields such as medicine, biology, and economics.

The expansion of regression methods also included nonparametric and robust regression techniques that relax some of the more restrictive assumptions of traditional parametric regression models, thereby providing greater flexibility in analyzing data that do not conform to normal distribution or that exhibit heteroskedasticity or non-linearity. These techniques are designed to be less sensitive to outliers and other forms of data that do not meet the assumptions of classical regression models. The use of bootstrapping and other resampling techniques has also improved the estimation of regression models by allowing for the estimation of the sampling distribution of almost any statistic.

A substantial advancement in the field was the development of the Generalized Additive Models (GAMs) by Hastie and Tibshirani in the 1980s, allowing for the modeling of non-linear relationships between the dependent and independent variables. GAMs can include both linear and non-linear terms and are particularly useful in situations

where the relationship between the variables is not well described by a straight line.

Furthermore, the latter part of the 20th century and the early 21st century have seen an explosion in the development of techniques for dealing with high-dimensional data—data that have a large number of variables relative to the number of observations. Techniques such as ridge regression, lasso regression, and elastic net were developed to handle multicollinearity and overfitting, common problems when dealing with high-dimensional data sets.

The concept of regression has also been extended to the field of machine learning. Techniques like support vector machines, random forests, and neural networks can be seen as extensions of regression. They are designed to capture complex patterns in the data, but at their core, they also seek to predict a dependent variable from a set of independent variables.

Today, regression analysis remains a vital tool across diverse fields, including finance, medicine, social sciences, and natural sciences. With the increasing availability of data and computational power, new regression techniques and applications are constantly emerging. Machine learning algorithms, for instance, often incorporate regression models as building blocks, further expanding the reach and impact of this powerful analytical framework.

1.2.2 Modern Interpretation of the Term Regression

The term "regression" has transcended its original biological connotations to denote a versatile, methodological tool used for prediction and inference regarding relationships among variables. It goes beyond simply finding the "best-fitting line" to exploring the mechanisms and dynamics driving the relationship. Linear models, though still widely used, are seen as a specific case within a broader spectrum. Non-linear, time-series, and other complex models are employed to better represent real-world phenomena.

Modern regression emphasizes constructing models that capture the relationship between variables based on underlying theoretical frameworks or empirical observations. It acknowledges the inherent uncertainty and potential biases in data and models. Statistical techniques like confidence intervals, hypothesis testing, and model selection aim to quantify and address these uncertainties, leading to more robust and reliable conclusions. Key features of modern regression are:

- **Wide range of models:** Linear, non-linear, multivariate, time series, and other specialized models are available.
- **Focus on causality:** While not always possible, modern regression techniques attempt to infer causal relationships between variables.
- **Robustness and diagnostics:** Techniques exist to address issues like outliers, multicollinearity, and heteroscedasticity, ensuring model reliability.
- **Computational efficiency:** With advancements in computing power, regression analysis can handle large datasets and complex models.

The modern interpretation of regression signifies its transformation from a biological concept to a powerful statistical and econometric tool. It allows us to model, quantify, and understand relationships between variables in diverse fields, enabling data-driven decision-making and uncovering valuable insights. Modern interpretations emphasize different aspects of this process:

1. **Prediction:** In contemporary usage, regression is primarily associated with the prediction of values of a dependent variable based on one or more independent variables. For example, in econometrics, regression might be used to predict consumer spending based on income and wealth levels. The coefficients obtained from regression models quantify the expected change in the dependent variable for a one-unit change in an independent variable, holding other variables constant.
2. **Inference:** Apart from prediction, regression is used for inference about the relationships between variables. This means testing hypotheses about these relationships, such as whether they are positive, negative, or non-existent. In econometrics, such inferences might involve understanding the impact of a policy change on economic growth, or the sensitivity of investment to interest rate changes.
3. **Estimating Relationships:** Regression models are used to estimate the functional form of relationships between variables. For instance, in econometrics, we might use regression to estimate the elasticity of demand for a product, or the marginal propensity to consume out of income.
4. **Causal Analysis:** Modern regression analysis, especially within econometrics, is often concerned with identifying causal relationships. This involves distinguishing correlation from causation and employing

techniques such as instrumental variables, difference-in-differences, or regression discontinuity designs to infer causality.

In the context of statistical analysis, especially econometrics, the modern interpretation of regression focuses on its role as a fundamental tool for quantitative analysis. It features are:

1. **Sophistication in Techniques:** Modern regression includes a variety of sophisticated techniques to handle diverse data structures and complexities. These range from simple linear regression to multiple regression, logistic regression for binary outcomes, Poisson regression for count data, and survival models for time-to-event data, among others.
2. **Accounting for Non-Linearity and Interaction:** Contemporary regression analysis can capture non-linear relationships through polynomial regression or the use of splines. It also accounts for interactions between variables to understand how the relationship between two variables may change at different levels of a third variable.
3. **Robust and Resistant Methods:** Modern regression methods are robust to violations of classical assumptions like normality of errors or homoscedasticity. Techniques such as robust regression are employed to provide reliable results even when outliers or other anomalies are present.
4. **Handling High-Dimensional Data:** With the advent of big data, regression techniques have adapted to handle high-dimensional datasets. Regularization methods like ridge regression, lasso, and elastic net have become important tools for variable selection and preventing overfitting.
5. **Model Selection and Validation:** Contemporary practice in regression involves rigorous model selection and validation techniques. Cross-validation, information criteria like AIC and BIC, and other methods are used to select the best model from a set of candidates and to assess the model's predictive power.
6. **Software and Computation:** Modern regression analysis is facilitated by sophisticated software that can handle complex models and large datasets. Statistical software packages now offer extensive libraries for regression analysis, making advanced techniques accessible to a wide range of users.

7. **Machine Learning and Artificial Intelligence:** Regression concepts have been extended into the realms of machine learning and artificial intelligence. Techniques such as regression trees, random forests, and neural networks represent an advanced form of regression used for both classification and prediction in complex datasets.

Modern interpretation of regression serves as the backbone for empirical research, guiding both theoretical exploration and practical policy implications. It is not merely a statistical technique but a comprehensive approach to understanding and interpreting the world through data. The modern view of regression is one of a dynamic, evolving discipline, integral to the field of econometrics and its application to economic data.

Today, regression analysis is an indispensable tool in the data scientist's toolkit. It continues to grow and adapt as new challenges in data analysis arise and exemplifies the dynamic nature of the field and its ability to adapt to the ever-changing landscape of data analysis. As we collect more data and develop new computational tools, it is likely that regression analysis will continue to evolve, providing ever more powerful tools for understanding the world around us.

1.2.3 Statistical Versus Deterministic Relationship

The world around us is full of connections and influences. Understanding these relationships is crucial in various fields, from science to economics. Two fundamental types of relationships govern these connections: statistical and deterministic. While both reveal linkages between variables, they differ significantly in their nature and predictability.

A statistical relationship is often probabilistic, meaning that it is based on tendencies or patterns observed within data rather than certainties. For example, a statistical relationship might show that taller people tend to have larger shoe sizes. This does not mean that every tall person has large feet, but there is a trend or tendency that can be observed in a population. Consider another example, the relationship between smoking and lung cancer. While smoking increases the risk of lung cancer, it does not guarantee it. Some smokers never develop the disease, while non-smokers sometimes do. This is because statistical relationships involve probabilities and trends, rather than perfect predictions. Knowing one variable (smoking) gives you a likelihood, not a certainty, of the other (lung cancer).

Statistical relationships are quantified using correlation coefficients or regression analysis, which measure the strength and direction of the association between variables. In the case of regression, the relationship also involves predicting the value of one variable based on the value of another. Statistical relationships are characterized by:

- **Probabilistic nature:** The influence of one variable on another is expressed through probabilities or tendencies.
- **Presence of randomness:** Chance or unknown factors play a role in the relationship.
- **Correlational, not causal:** The observed connection doesn't necessarily imply causation.

Statistical relationship suggests that a change in one variable is associated with a change in another, but this association is not necessarily one of direct causation and may include randomness or variability.

A deterministic relationship, on the other hand, is one where a certain input will always produce the same output, without randomness or variation involved. This type of relationship can be described by an exact mathematical equation, and given the same starting conditions, the outcome will always be the same. An example of a deterministic relationship is the calculation of the circumference of a circle from its radius. The equation (where is the circumference and is the radius) will always yield the exact circumference if the radius is known. There is no variability or uncertainty in this calculation; it is a fixed, predictable relationship.

In a deterministic relationship, knowing the value of one variable allows you to exactly calculate the value of the other. This perfect predictability stems from the underlying physical or mathematical laws governing the relationship. For instance, knowing the mass and acceleration of an object lets you determine its exact velocity using the formula . Similarly, in Ohm's Law, measuring the voltage across a resistor allows you to precisely calculate the current flowing through it. Deterministic relationships are often characterized by:

- **Perfect predictability:** Knowing one variable guarantees the exact value of the other.
- **Absence of randomness:** No element of chance influences the relationship.

- **Clear functional dependence:** A precise formula or equation describes the relationship.

In summary, while statistical relationships account for trends, probabilities, and the influence of various factors, deterministic relationships rely on fixed, predictable formulas. The key differences between the two can be summarized as:

Feature	Deterministic Relationship	Statistical Relationship
Prediction	Exact and guaranteed	Estimate based on probabilities
Randomness	None; exact relationships	Chance and uncertainty play significant role in the relationship between variables
Causality	Relationships imply causation	Relationships only show correlation; causation needs to be ascertained through statistical tests
Examples	Laws of Physics; Engineering calculations	Exam scores and study hours; Economic trends

Understanding the difference between these relationships is crucial for interpreting data and drawing accurate conclusions. Deterministic relationships allow for precise predictions, while statistical relationships provide insights into trends and probabilities. The world is often complex and nuanced, and both types of relationships play vital roles in our understanding of it.

1.2.4 Regression Versus Causation

Understanding the distinction between regression and causation is crucial in the empirical examination of data across various disciplines. Conflating correlation, revealed by regression, with causation can lead to pitfall in analysis and interpretation of results. While regression can unveil statistical associations, the leap to inferring causal connections requires an extra layer of scrutiny.

Regression analysis, in its various forms, determines the degree to which a dependent variable changes in response to an independent variable. Specifically, it establishes a statistical association quantifying the dependent variable's (Y) change in response to variations in the independent variable(s) (X). However, the fundamental premise of regression is correlation, which measures the strength and direction of a relationship between variables, without necessarily implying a cause-and-effect relationship. For instance, a regression analysis could reveal how sales might vary with changes in advertising expenditure. This association is expressed through a regression line or equation, capturing the trend but not necessarily the underlying mechanism.

The crux of the issue lies in the inherent inability of regression to establish causality. While a strong association between variables might exist, it is crucial to remember that other factors, unaccounted for in the model, could be the true drivers of the observed relationship. Consider the classic example of ice cream sales and drowning deaths. Regression might show a positive correlation, but the true cause of drownings likely lies in factors like warmer weather, increased pool usage, and not the consumption of ice cream.

Several factors can confound the causal interpretation of regression results. Reverse causality occurs when the dependent variable, instead, influences the independent variable. For instance, studying the link between stress and smoking might reveal a positive association, but it could be that individuals experiencing stress turn to smoking as a coping mechanism, reversing the causal direction. Omitted variable bias arises when pertinent variables influencing both X and Y are excluded from the analysis. This creates a misleading association between X and Y, attributing to X an effect truly stemming from the omitted variable.

The quintessence of regression lies in prediction. It establishes a model that, given the value of an independent variable, can predict the average outcome for the dependent variable. This predictive capability is immensely valuable in many fields, including economics, where it is employed to forecast variables like inflation, growth, or consumption based on observed historical data. However, the regression model is inherently limited to the data it analyzes and is influenced by the scope of variables included in the model.

Correlation, as famously stated, does not imply causation. To move from correlation to causation, additional evidence is necessary. While regression alone cannot definitively establish causality, several approaches can strengthen causal inference. This evidence may come from randomized controlled trials, longitudinal studies, or the use of statistical techniques that can control for confounding variables, such as instrumental variable analysis. In econometrics, for instance, researchers often use such methods to infer causal relationships, such as the impact of education on earnings, by controlling for factors like ability, family background, and socio-economic status.

The gold standard for establishing causation is the randomized controlled trial, where subjects are randomly assigned to different conditions. This randomization ensures that both known and unknown factors are equally distributed across groups, isolating the effect of the independent variable on the dependent variable. However, in many situations, such experiments are either unethical or impractical. Therefore, researchers rely on observational data, employing statistical models and methods to approximate experimental conditions and infer causal relationships.

Quasi-experimental designs mimic RCTs in spirit, leveraging naturally occurring situations that resemble random assignment. Observational studies can be enhanced by employing instrumental variables, which influence the independent variable but are not directly associated with the dependent variable, potentially mitigating confounding factors. Finally, statistical techniques like propensity score matching aim to create comparable groups between which causal effects can be estimated.

On the other hand, causation delves deeper into the dynamics between variables, asserting a cause-and-effect relationship. When one asserts that variable A causes variable B, it implies that changes in A directly bring about changes in B. Establishing causation requires rigorous experimental or observational methods that go beyond mere statistical associations. It often necessitates the control of extraneous variables that could influence the outcome, something that is typically addressed in experimental research designs but can be more challenging in observational studies.

The distinction between regression and causation is often blurred in empirical research. A strong correlation between two variables does not confirm that one variable causes the other. For example, ice cream

sales and drowning incidents may be strongly correlated because both are higher in the summer months, but it would be fallacious to assert that ice cream consumption causes drowning incidents. Mistaking correlation for causation can lead to erroneous interpretations and misguided policies. Recognizing the limitations of regression and employing appropriate techniques to strengthen causal inference are crucial for extracting meaningful insights from the intricate tapestry of data. Remember, correlation paints a suggestive picture, but only by carefully disentangling its threads can we unveil the true causal forces at play.

1.2.5 Regression Versus Correlation

In statistical analysis, the concepts of regression and correlation are both pivotal and frequently juxtaposed. While often used interchangeably, they possess distinct purposes and offer unique insights. Their distinction, while subtle, has profound implications for the interpretation of data and the inferences drawn from statistical models. This essay endeavors to dissect these concepts, demarcate their distinctions, and articulate their respective applications within the domain of statistical analysis.

Correlation quantifies the strength and direction of a linear relationship between two quantitative variables--extent to which two variables fluctuate together. It produces a single value, the coefficient of correlation, denoted as 'r', ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear association. Correlation paints a broad picture, revealing whether variables move together, but it does not explain how one influences the other. Correlation is symmetric: the correlation between X and Y is the same as the correlation between Y and X. It is agnostic to the causal direction of the relationship, making it susceptible to misinterpretations. It does not differentiate between dependent and independent variables, merely reflecting the degree to which they co-vary. For instance, a high correlation between ice cream consumption and drowning rates might not imply one causes the other; both could be influenced by a hotter summer season. Take another example, the correlation between study hours and exam scores. A strong positive correlation indicates that, on average, students who dedicate more time tend to score higher. However, correlation does not imply causation. Other factors, like innate ability, could also play a role, and the relationship might not be perfectly linear.

Regression, on the other hand, goes beyond mere association. It models the relationship between a dependent variable (predicted) and one or more independent variables (predictors) using an equation. Returning to our study hours example, regression would provide an equation that estimates a student's likely exam score based on their study hours. The simple regression equation is y = mx + c, where 'y' is the dependent variable, 'x' is the independent variable, 'm' represents the slope of the line (indicating the rate of change in y for a one-unit change in x), and 'c' is the y-intercept. This equation allows us to predict the value of the dependent variable for a given value of the independent variable(s). The key advantage of regression lies in its ability to make predictions for unobserved data points, too. By using the fitted equation, we can estimate the value of the dependent variable for new data points. This allows us to explore "what-if" scenarios and gain insights into how changes in one variable might affect the other.

While correlation is bidirectional, regression is unidirectional: it aims to predict the value of the dependent variable based on the known values of the independent variable(s). Regression provides a causal interpretation, assuming all relevant factors are included in the model. However, this interpretation relies on strong theoretical justification and careful control for confounding variables, which can be challenging to achieve. Furthermore, regression is sensitive to outliers and model misspecification, potentially leading to inaccurate predictions.

The coefficient of determination, denoted as R^2, emerges from regression analysis and provides a measure of how well observed outcomes are replicated by the model. It is the square of the correlation coefficient when dealing with simple linear regression and serves to indicate the proportion of the variance in the dependent variable that is predictable from the independent variable.

The choice between correlation and regression hinges on the research question. Correlation is ideal for exploratory analysis, identifying potential associations and generating hypotheses. It is also useful for comparing the strength of relationships across different groups or variables. Correlation is a prerequisite for linear regression; regression assumes that there is a correlation to be described and modeled. However, a significant correlation does not guarantee a meaningful or significant regression equation. Correlation does not account for causality or the dependency

between variables, whereas regression is explicitly designed to model the potential causality from the independent to the dependent variable.

Regression shines when we want to predict the dependent variable from the independent variable(s) or explain the mechanism underlying the relationship. It is crucial for testing of hypotheses and building models. Moreover, regression analysis allows for the inclusion of multiple independent variables, facilitating a multifaceted view of how several predictors jointly impact the dependent variable. There are various regression techniques, each with its own strengths and weaknesses. Linear regression, the most common, assumes a linear relationship between the variables. More complex models, like polynomial regression, can capture non-linear relationships, but require careful interpretation and consideration of overfitting. This is not the case with correlation, which is confined to the assessment of a single bivariate relationship at a time. However, the line between correlation and regression can blur. Correlation can be a steppingstone to regression, informing which variables to include in the model. Additionally, some regression techniques, like partial correlations, aim to isolate the relationship between two variables while controlling for others, mirroring the purpose of correlation analysis.

In practical applications, correlation is often a preliminary step, providing a quick insight into the possibility of a relationship between variables. Regression is the subsequent, more sophisticated step that models this relationship and can be used to make predictions or inferences. For instance, in finance, correlation might be used to identify assets that move together for portfolio diversification, while regression could be employed to predict future asset prices based on various economic indicators. In econometrics, regression is indispensable. Econometricians routinely employ regression models to estimate the impact of policy changes or economic conditions on various socioeconomic outcomes. The interpretation of regression coefficients requires careful consideration of the underlying theory and the context of the data. The key differences between the two can be summarized as follows:

Feature	Correlation	Regression
Directionality	Agnostic to the direction of influence between variables	Assumes a clear dependent and independent variable(s)
Prediction	Cannot predict the value of one variable based on another	Builds a predictive model
Model Building	Does not involve model building	Constructs an equation to represent the relationship
Causality	Does not imply causation	Attempts to estimate the causal effect of one variable on another

Correlation and regression are not rivals, but complementary tools. Correlation offers a quick assessment of association. It is also useful when the relationship is not of primary interest, but understanding the presence or absence of an association is important whereas regression delves deeper into the model and predict. Regression, however, is the tool of choice when prediction or understanding the impact of one variable on another is the primary objective. It allows for more nuanced analysis, quantifying the strength of the relationship and enabling predictions for new data points.

Understanding their unique strengths and limitations is paramount for drawing accurate conclusions from data and making informed decisions based on statistical evidence. It is important to acknowledge the limitations inherent to both regression and correlation. Correlation's inability to imply causation is well-noted; it is silent on whether one variable influences the other. Regression can suggest causality when used with a theoretical framework and empirical evidence that supports a causal link. However, without proper experimental or quasi-experimental design, regression too can fall prey to the fallacy of assuming causation from mere association.

While correlation and regression are interrelated, their roles in statistical analysis are distinct. Correlation provides a measure of the linear relationship between variables without implying causation, serving as a foundational stepping-stone to more intricate analyses. Regression, on the other hand, builds on this foundation to model the direction and strength of relationships, allowing for

prediction and the possibility of causal inference within the constraints of the data and accompanying assumptions. Both are formidable tools in the statistician's arsenal, each with its specific utility, limitations, and interpretive nuances. Understanding and applying these concepts with precision is paramount for the rigorous analysis and interpretation of quantitative data.

1.2.6 Terminology and Notation

Econometrics employs a specific language of symbols and terms. Understanding these is crucial for analysis and interpreting results. Following is the list of key terminologies and notations commonly used in elementary econometrics:

A. **Population and Sample**

   a. Population (**N**): Refers to the entire group that one wants to draw conclusions about.

   b. Sample (**n**): A subset of the population that is used to infer conclusions about the population.

   c. Population Mean (μ): Represents the mean of values in the population.

   d. Population Standard Deviation (σ): Represents the standard deviation of values in the population.

B. **Variables**

   a. Dependent Variable (**Y**): The outcome or the variable of interest that the model aims to predict or explain (e.g., wage, price).

   b. Independent Variables ($\mathbf{X_1, X_2, ..., X_k}$): Factors that are posited to have an effect on the dependent variable (e.g., education level, experience). These are also called explanatory variables or predictors.

   c. Control Variables (**Z**): Additional independent variables included to account for other factors affecting Y, denoted by Z (e.g., age, gender).

C. **Parameters**

   a. Intercept (α): The constant term in the regression equation, representing the value of Y when all X's are zero.

   b. Slope Coefficient (β): The coefficient of each independent variable, measuring the change in Y for a one-unit increase in X, holding all other variables constant.

D. **Coefficients**

a. Beta Coefficients ($\boldsymbol{\beta_0, \beta_1, ..., \beta_k}$): Parameters in a population regression line that represent the effect of each independent variable on the dependent variable.

b. Estimated Coefficients ($\mathbf{b_0, b_1, ..., b_k}$): Sample estimates of the population beta coefficients, obtained through regression analysis.

E. **Error Term**

a. Population Error Term (ε): The error in the prediction of the dependent variable in the population regression, capturing the effect of all other variables not included in the model.

b. Residual (**e**): The difference between the observed value and the predicted value of the dependent variable in the sample.

F. **Regression Equation**

a. Simple Linear Regression: $Y = \alpha + \beta X + \varepsilon$ (one independent variable)

b. Multiple Linear Regression: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$ (more than one independent variables)

c. Population Regression Line: ($Y = \beta^0 + \beta^1 X^1 + ... + \beta_k X_k + \varepsilon$ (unobserved, unknown).

d. Sample Regression Line: $\hat{Y} = b_0 + b_1 X_1 + ... + b_k X_k$ (estimation of population regression line)

G. **Estimation Methods**

a. Ordinary Least Squares (**OLS**): The method of estimating the unknown parameters in a linear regression model by minimizing the sum of squared residuals.

b. Maximum Likelihood Estimation (**MLE**): A method of estimating parameters by maximizing the likelihood function.

H. **Assumptions**

a. Classical Linear Regression Model (**CLRM**) Assumptions: A set of assumptions required for OLS estimators to be **BLUE** (Best Linear Unbiased Estimators). These include linearity, no perfect multicollinearity, exogeneity, homoskedasticity, and normality of errors.

I. **Goodness-of-Fit Measures**

a. R-squared (**R^2**): The proportion of the variance in the dependent variable that is predictable from the independent variables. Ranges from 0 to 1, with higher values indicating better fit.

b. Adjusted R-squared: Modification of R^2 that adjusts for the number of explanatory variables in a model relative to the number of data points.

c. Standard Error (**SE**): Measures the variability of the estimated coefficient around its true value. Lower SE indicates more precise estimates.

J. **Test Statistics**

a. *t*-statistic(***t***): Used to test if a single parameter is significantly different from zero. Compares the estimated coefficient to its standard error, assessing its statistical significance. A t-statistic greater than 2 (in absolute value) usually indicates a statistically significant coefficient.

b. *F*-statistic (***F***): Used to test if a group of parameters are jointly different from zero. Tests the overall significance of the regression model, assessing whether the independent variables jointly explain a statistically significant portion of the variance in Y.

K. **Hypothesis Testing**

a. Null Hypothesis (**H_0**): A statement that there is no effect or no relationship (e.g., $\beta_1 = 0$).

b. Alternative Hypothesis (**H_1**): The statement that there is an effect or a relationship (e.g., $\beta_1 \neq 0$).

L. **Statistical Significance**

a. Confidence Interval: A range of values derived from the sample statistics that is likely to cover the true population parameter.

b. ***p***-value: The probability of obtaining the observed sample results when the null hypothesis is true. A small p-value (typically $\leq$ 0.05) indicates strong evidence against the null hypothesis.

M. **Diagnostics**

a. Multicollinearity: A situation in which two or more explanatory variables in a multiple regression model are highly linearly related.

b. Heteroskedasticity: The condition in which the variance of the residuals is not the same across all levels of the independent variables.

c. Autocorrelation: Correlation of the error term over time. High correlation among independent variables, leading to imprecise estimates.

d. Stationarity: A characteristic of a time series in which the properties do not depend on the time at which the series is observed.

N. **Visualizations**

a. Scatter Plot: Displays the relationship between two variables, with each data point representing a single observation.

b. Regression Line: A line drawn through the scatter plot, representing the predicted values of Y based on the regression equation.

For example, in a theoretical model hypothesizing a linear relationship between wages earned (outcome variable) and education and experience (predictor variables), the regression model can be formulated as:

$$\text{Wages} = \alpha + \beta_1 (\text{Education}) + \beta_2 (\text{Experience}) + \varepsilon$$

Running an ordinary least squares procedure on this model will give estimations of population parameters interpreted as follows:

α = expected wage when both Education and Experience are zero

β_1 average increase in wages for each additional year of education, holding Experience constant

β_2 average increase in wages for each additional year of experience, holding Education constant

Data needed for estimating the relationship can either be collected by the researcher for the specific purpose of conducting this study (primary data) or can be extracted from datasets collected by other researchers or institutions for their own purposes (secondary data).

1.2.7 The Nature and Sources of Data for Econometric Analysis

Econometric analysis hinges on the quality and nature of its data. Understanding the data landscape is critical for researchers, as it shapes

the types of questions we can ask, the models we can build, and the conclusions we can draw. The specific types of data and considerations may vary depending on the research area and chosen methodology. There are many ways in which data can be classified. Following are sone of the major classification of data types:

A. **Primary vs. Secondary Data**

a. **Primary data:** Collected directly by the researcher for a specific research question. Offers greater control over data quality and content, but can be expensive and time-consuming to acquire. Examples: Surveys, experiments, field observations.

b. Secondary data: Already collected and compiled by other organizations (government agencies, research institutions, private companies). Offers readily available data, often covering large populations and extended time periods, but may have limitations in terms of scope, quality, and control. Examples: Census data, financial databases, market research reports.

B. **Observational vs. Experimental Data**

a. **Observational Data:** Drawn from real-world observations, it is non-experimental, meaning researchers cannot manipulate variables directly. Examples include:

i. **Government data:** Census data, labor market statistics, trade data (Strengths: Large scale, often publicly available; Limitations: Potential biases, limited control over variables)

ii. **Surveys:** Household surveys, business surveys (Strengths: Rich information on individual/firm behavior; Limitations: Sample selection bias, measurement error)

iii. **Financial data:** Stock prices, exchange rates (Strengths: High frequency, readily available; Limitations: Endogeneity, market microstructure issues)

b. **Experimental Data:** Derived from controlled experiments where researchers manipulate variables to observe their causal effects. While rare in economics due to cost and ethical concerns, it offers valuable insights when feasible. (Strengths: Direct identification of causal effects; Limitations: Limited generalizability, artificiality of lab settings)

C. **Cross-sectional vs. Time-series vs. Panel Data**

a. Cross-Sectional Data: Refers to data collected on different entities at a single point in time or over a very short period. These entities could be individuals, households, firms, countries, or a variety of other units. In a cross-sectional dataset, each observation represents a single entity and can include variables of interest such as income, education level, or expenditure on healthcare. The main advantage of cross-sectional data is that it can provide a snapshot of the population at a specific point in time, allowing for comparisons across different entities. However, one of the challenges with cross-sectional data is the potential for omitted variable bias, where unobserved differences between the entities can influence the results of the econometric analysis.

b. Time-Series Data: Consists of observations on a variable or several variables over time. These are sequential measurements taken at regular or irregular intervals. For example, time-series data can track the GDP of a country, stock prices, interest rates, or unemployment rates across several years or quarters. The analysis of time-series data allows economists to understand temporal dynamics, identify trends, and make forecasts. However, time-series data can be complicated to analyze due to potential issues such as autocorrelation (where the value of a variable is correlated with its own past values), non-stationarity (where properties of the series like mean and variance change over time), and seasonality (periodic fluctuations due to seasonal factors).

c. Panel Data: Also called longitudinal data, combines the features of cross-sectional and time-series data. It consists of observations on multiple entities over multiple time periods. This data type allows for richer modeling of behavior because it captures both the inter-temporal dynamics for each entity and the differences between entities. Panel data provide several advantages for econometric analysis. They can improve the estimation efficiency, control for individual heterogeneity, and provide a more complex structure that allows for more varied econometric models. However, panel data also come with challenges, such as the complexity of the models required to analyze them, potential for missing data issues, and the need for advanced techniques to handle the interdependence of observations within an entity over time.

In addition to these primary data types, econometricians must also consider the sources and quality of data. Data can be obtained from surveys, administrative records, experiments, or observational studies. Each source has its strengths and weaknesses that can affect the reliability and validity of econometric inferences. Furthermore, issues such as measurement errors, sampling bias, and data cleaning can significantly influence the analysis.

The data should be free from errors and inconsistencies. Careful data collection and cleaning are essential. Data cleaning and pre-processing are essential steps to ensure the reliability of analysis. When data does not accurately reflect the underlying variable, leading to biased estimates. For example, self-reported income in surveys may be underestimated. Missing values, outliers, and inconsistencies need careful attention. Missing data can introduce bias and limit the scope of analysis. Imputation techniques may be necessary to address missing values. The level of detail in the data determines the level of analysis possible. More detailed data allows for more nuanced insights but may also increase complexity and computational demands. And, finally, the data should represent the population of interest accurately. Selection bias can occur if specific groups are over- or underrepresented. When the sample of data does not represent the population of interest, leading to biased estimates. For example, A survey on health outcomes may only include people who are already interested in health.

Econometric studies make use of various types of data, each serving different research purposes and providing unique insights into economic behaviors and trends. The choice of data source for an econometric study depends heavily on the specific research question being investigated. However, some commonly used data sources for econometric studies include:

A. **Government Agencies**

   i. National Statistical Offices: These agencies collect and publish a wide range of data on various topics, including demographics, economics, labor, and social indicators. Examples include:

      a. United States: U.S. Census Bureau: [URL https://www.census.gov/]

      b. India: Ministry of Statistics and Programme Implementation: [URL https://www.mospi.gov.in/]

c. The World Bank: World Development Indicators: [URL https://data.worldbank.org/]

ii. Central Banks: Central banks collect and publish data on monetary policy, financial markets, and the economy. Examples include:

a. U.S. Federal Reserve: Board of Governors of the Federal Reserve System [URL https://www.federalreserve.gov/]

b. Reserve Bank of India: [URL https://rbi.org.in/]

iii. Other Government Agencies: Many other government agencies collect and publish data relevant to specific sectors or topics. Examples include:

a. U.S. Bureau of Labor Statistics: Bureau of Labor Statistics: [URL https://www.bls.gov/]

b. India's National Sample Survey Organization: [URL https://www.india.gov.in/nsso-reports-publications]

B. **International Organizations**

i. International Monetary Fund (IMF): The IMF collects and publishes data on global economic indicators, including GDP, trade, and financial markets. [URL https://www.imf.org/]

ii. World Bank: The World Bank collects and publishes data on a wide range of topics related to development, poverty, and inequality. [URL https://data.worldbank.org/]

iii. Organization for Economic Co-operation and Development (OECD): The OECD collects and publishes data on economic, social, and environmental indicators for its member countries. [URL https://www.oecd.org/]

C. **Private Sector**

i. Commercial Databases: Many private companies provide access to economic and financial data through subscription-based services. Examples include:

a. Bloomberg Terminal: [URL https://www.bloomberg.com/professional/]

b. Thomson Reuters Datastream: [URL https://www.lseg.com/en/data-analytics/products/datastream-macroeconomic-analysis]

c. Worldscope: Harvard Business School, Baker Library [URL https://www.library.hbs.edu/find/databases/worldscope]

ii. Research Institutes: Some research institutes collect and publish data on specific topics. Examples include:
   a. National Bureau of Economic Research (NBER): [URL https://www.nber.org/]
   b. Pew Research Center: [URL https://www.pewresearch.org/]
iii. Individual Surveys: Researchers may also collect their own data through surveys or experiments.

Some of the most commonly used data repositories are:

A. **Global**

i. World Bank Open Data: A comprehensive repository of data on development indicators from the World Bank. [URL https://data.worldbank.org/]
ii. United Nations Data: A repository of data from various UN agencies on a wide range of topics. [URL https://data.un.org/]
iii. The World Bank Indicators: A collection of indicators on various topics related to development. [URL https://databank.worldbank.org/]
iv. OECD iLibrary: A repository of data and publications from the OECD. [URL https://stats.oecd.org/]

B. **India**

i. India Open Data Platform: A government initiative to promote open access to data from various ministries and departments. [URL https://data.gov.in/]
ii. National Data Repository: A repository of data from various government agencies in India. [URL https://www.ndsindia.org/]
iii. MOSPI Data Portal: A data portal from the Ministry of Statistics and Programme Implementation. [URL https://mospi.gov.in/data]
iv. Reserve Bank of India Database on Indian Economy (RBI-DEIE): A database of economic and financial data from the Reserve Bank of India. [URL https://cimsdbie.rbi.org.in/DBIE/]

These are just a few examples, and the best data source for any particular study will depend on the specific research question and needs.

1.2.8 Summary

Econometrics is a pivotal field that bridges economics with statistical methods to analyze economic data, enabling economists to test hypotheses and forecast future trends. This discipline's core revolves around regression analysis, a statistical technique used for estimating the relationships among variables. Over time, econometrics has evolved, incorporating complex models and methods to address various challenges, including non-linear relationships, endogeneity, and heteroscedasticity.

Econometrics has, as a discipline, transformed from simple linear regression models to more sophisticated econometric techniques. These advancements have been crucial for understanding economic phenomena in a more nuanced manner, allowing for more accurate predictions and policy evaluations. Distinctions between statistical and deterministic relationships, regression versus causation, and regression versus correlation are essential for interpreting econometric results correctly, as they highlight the limitations and potential misinterpretations of statistical analyses.

Econometric analysis relies on various data types, including cross-sectional, time-series, and panel data. Each data type offers unique insights but also presents specific challenges that econometricians must navigate. Modern econometrics has expanded to include techniques like machine learning and big data analytics, reflecting the field's adaptability to new technological advancements. This evolution signifies a shift towards more dynamic and flexible approaches to economic analysis, capable of handling the complexity and volume of contemporary data.

1.2.9 Keywords

Autocorrelation: The correlation of a variable with itself across different time intervals. In time-series data, it refers to the similarity between observations as a function of the time lag between them, which can affect the assumptions of classical regression models.

Causation vs. Correlation: A fundamental distinction in statistical analysis where causation implies a cause-effect relationship between variables, whereas correlation indicates a mutual relationship without implying cause and effect.

Cross-sectional Data: Data collected at a single point in time across several subjects, providing a snapshot of a system or phenomena.

Deterministic Relationship: A relationship where one variable deterministically determines another, without randomness in the association.

Endogeneity: A condition in regression analysis where an explanatory variable is correlated with the error term, leading to biased and inconsistent estimates.

Heteroscedasticity: A situation in regression models where the variance of the error terms varies across observations, potentially leading to inefficient estimates.

Multicollinearity: A situation in regression analysis where two or more independent variables are highly correlated, making it difficult to distinguish their individual effects on the dependent variable.

Panel Data: Data that combines cross-sectional and time-series data, tracking the same subjects across multiple time periods, enabling more nuanced analyses of changes over time.

Regression Analysis: A statistical technique used to estimate the relationships among variables. It is fundamental in econometrics for modeling and analyzing several variables' interactions.

Statistical Relationship: A relationship between variables that is influenced by chance, implying that the relationship can vary and is not fixed or deterministic.

Time-Series Data: Data collected over several time periods, allowing for the analysis of trends, cycles, and other temporal effects on the subject of interest.

1.2.10 Self-assessment Questions

1. What is the historical origin of the term "regression" in statistics?
2. How does the modern interpretation of regression differ from the concept of "regression towards the mean"?
3. Distinguish between statistical and deterministic relationships in the context of regression analysis.
4. Explain the difference between correlation and causation, and how regression analysis addresses this distinction.
5. Describe the key terminology used in regression analysis, including dependent and independent variables, regression line, intercept, slope, and residuals.

6. What are the different types of data commonly used in econometric analysis?
7. Discuss the importance of data quality in regression analysis and the potential consequences of using poor-quality data.
8. Can you think of an example where regression analysis has been used in a field other than econometrics?
9. What are some of the limitations of using regression analysis to draw conclusions about causal relationships?
10. How can you ensure that your interpretation of regression results is statistically sound and avoids common pitfalls?

1.2.11 References

1. **Introductory Econometrics: A Modern Approach** by Jeffrey M. Wooldridge. This textbook is renowned for its clarity and practical approach to econometrics. It covers basic concepts, regression analysis, and more advanced topics like panel data and instrumental variables. Wooldridge's book is well-suited for undergraduates and graduate students alike, offering intuitive explanations and real-world examples.
2. **Econometric Analysis** by William H. Greene. Greene's book is a comprehensive guide that delves into both the theory and application of econometrics. It's known for its rigorous treatment of statistical methods and its detailed exploration of econometric models, making it a staple for graduate students and professionals seeking a deeper understanding of the field.
3. **Econometrics** by Fumio Hayashi. This book provides a solid foundation in the principles of econometrics, focusing on estimation, inference, and other fundamental concepts. Hayashi emphasizes the use of matrix algebra in understanding econometric models, making it suitable for students with a strong quantitative background.
4. **Econometrics** by Bruce D. Hansen. This textbook offers a more concise and theoretical treatment of econometrics, suitable for undergraduates with a strong grasp of statistics and mathematics.
5. **Using Econometrics: A Gentle Introduction** by A. Colin Cameron and Pravin K. Trivedi. This textbook focuses on the practical application of econometrics, presenting various econometric techniques through real-world examples.

UNIT - II : Two Variable and Multiple Regression Analysis

Lesson 2.1 – Simple Linear Regression Model

Structure

2.1.1 Introduction to Regression Analysis

Regression analysis is a fundamental statistical technique used to investigate the relationships between variables. At its core, regression analysis allows us to model, examine, and predict the association between a dependent variable and one or more independent variables. The dependent variable, also known as the outcome or response variable, is the variable we are trying to understand or predict. The independent variables, often termed predictors or explanatory variables, are the variables presumed to influence the dependent variable. The importance of regression analysis lies in its versatility; it can be applied to a myriad of economic questions, from assessing the impact of education on earnings to understanding the determinants of consumer spending.

The simplest forms of regression analysis is the linear regression model, which assumes a linear relationship between the dependent and independent variables. This model is instrumental in scenarios where the objective is to understand how a unit change in the independent variable affects the dependent variable. For instance, consider the relationship

between household income (independent variable) and expenditure on consumer durables (dependent variable).

The relationship is called linear because the datapoints of income and expenditure can be modeled as a geometrical straight line with parameters, represented as coefficients of the straight line:

$$y=c+mx$$

A linear regression model could help quantify how changes in income influence spending on consumer durables, offering valuable insights for businesses and policymakers alike. However, the real world is seldom linear or simple. Economic relationships often involve multiple factors interacting in complex ways, necessitating the use of multiple and nonlinear regression analysis.

The multiple regression model incorporates several independent variables to explain the variation in the dependent variable, providing a more nuanced understanding of the economic dynamics at play. For example, to analyze the factors influencing house prices, a multiple regression model might include variables such as location, size, age, and proximity to amenities. Such an analysis can reveal the relative importance of these factors, guiding both buyers in their decisions and governments in their housing policies.

Beyond linear models, regression analysis encompasses a variety of techniques to tackle different types of data and relationships. Logistic regression, for instance, is used when the dependent variable is binary, such as when studying the likelihood of a consumer defaulting on a loan based on their credit score and income level. Similarly, time series regression models are crucial for analyzing data that is sequential in time, enabling economists to forecast future economic conditions based on past trends.

The application of regression analysis extends across the economic spectrum; it has seemingly endless applications–to illustrate a few:

- Sales Forecasting: A business might use regression to predict sales based on factors like advertising expenditure, historical sales data, and economic indicators.
- Medical Diagnosis: Regression can help determine risk factors for a disease. For instance, modeling blood pressure based on age, weight, diet, and exercise.

- Stock Market Modeling: Stock prices can be modeled using regression analysis with variables such as company earnings, interest rates, and market sentiment.
- House Price Prediction: Regression helps predict housing prices based on characteristics like square footage, number of bedrooms, location, and neighborhood amenities.
- Student Performance Analysis: Educators may use regression to analyze student performance based on factors such as prior grades, study habits, and socioeconomic background.

Despite its widespread application and utility, regression analysis is not without limitations. The accuracy of regression models depends heavily on the quality of the data and the appropriateness of the chosen model. Issues such as multicollinearity among independent variables, heteroscedasticity, and autocorrelation can undermine the reliability of the estimates and lead to misleading conclusions. For linear regression to produce reliable results, certain assumptions must hold true:

- Linearity: The relationship between the dependent and independent variable(s) is linear.
- Independence of Errors: Errors are independent of each other (no autocorrelation).
- Homoscedasticity: Constant variance of errors across different values of independent variables.
- Normality of Errors: Errors are normally distributed.

The choice of model, the accuracy of data, and the assumptions underlying the statistical methods are paramount considerations that determine the reliability and validity of the analysis. The process of conducting a regression analysis involves several critical steps:

- Data Preparation: Ensure data is clean, formatted correctly, and missing values are addressed. Outliers should be examined.
- Exploratory Data Analysis (EDA): Visualize using scatterplots, histograms, and other techniques to check assumptions and spot patterns.
- Model Fitting: Select the appropriate regression type and use statistical software to estimate the model coefficients.
- Model Evaluation: Assess the model fit using R^2, residuals, and significance tests.

- Diagnostics: Check for violations of assumptions (e.g., linearity, homoscedasticity) and address them if necessary.
- Prediction and Interpretation: Utilize the model to predict values of the dependent variable for new data points and carefully interpret results.

The economists must exercise caution, employing diagnostic tests and robust statistical techniques to validate their models and ensure the integrity of their findings.

2.1.2 Economic Versus Statistical Model

In econometrics, economic and statistical models serve as indispensable tools for analyzing and comprehending complex economic phenomena. While economic models provide a theoretical framework for economic relationships, statistical models offer the means to quantify and test these relationships using empirical data.

An economic model is an abstract, simplified representation of economic processes or behaviors. It employs a set of variables and logical or quantitative relationships to depict how these variables interact within an economic system. Economic models are built upon theoretical assumptions and aim to capture the essence of economic decision-making and outcomes. For instance, the supply and demand model in microeconomics simplifies the market mechanism by assuming ceteris paribus (all other factors being constant), focusing on the relationship between the price of a good and the quantity supplied and demanded. Key features of an economic model include:

- **Variables:** Economic models incorporate both endogenous variables (determined within the model) and exogenous variables (determined outside the model). For example, a model analyzing consumer demand might include price and income as exogenous variables, while the quantity demanded becomes the endogenous variable.
- **Assumptions:** These form the foundation of economic models and often involve simplifying complex realities to make the model more tractable. Assumptions like perfect information, rational economic agents, or the absence of market frictions are common.
- **Relationships:** The core of the model lies in the postulated relationships between the variables. They could be functional

forms (i.e., linear, non-linear), causal directions, and qualitative expectations about the interplay of variables. For instance, the law of demand suggests an inverse relationship between price and quantity demanded.

A classic example of an economic model is the demand model. In its basic form, it theorizes that:

- The quantity demanded of a good/service is inversely related to its price.
- Consumer's income influences demand (usually positively).
- The price of related goods (substitutes or complements) also impacts demand.

This model uses assumptions like consumers acting rationally to maximize satisfaction within their budget constraints.

In contrast to the theoretical nature of economic models, statistical models focus on the empirical side of analysis. They utilize statistical techniques to establish relationships between variables based on real-world observations. Statistical models help estimate the strength and direction of the relationships hypothesized by economic models and provide a means to test them. By incorporating probability distributions, statistical models account for uncertainty and variability in data, facilitating inference about the broader population from sample observations. A simple linear regression model, for example, can be used to understand how changes in one variable (independent variable) are associated with changes in another (dependent variable), incorporating randomness and uncertainty inherent in real-world data. A statistical model generally entails:

- **Data:** The backbone of a statistical model is a dataset containing observations on the relevant variables identified by the economic model.
- **Functional Form:** This mirrors the hypothesized relationship; often, it's the mathematical equation drawn from the economic model.
- **Stochastic component:** Statistical models recognize that the real world is rarely as precise as theoretical models. It includes a random error term to account for unexplained variation and other factors not explicitly included in the model.

The demand model, previously explained from a theoretical perspective, can be converted into a statistical representation:

$$Q_d = \beta_0 + \beta_1 P + \beta_2 I + \beta_3 P_r + \varepsilon$$

Where:

- Q_d = Quantity demanded
- P = Price of the good
- I = Consumer income
- P_r = Price of a related good
- β_0, β_1, β_2, β_3 = Model parameters (coefficients) to be estimated
- ε = Random error term

The conversion of an economic model into a statistical model is driven by the necessity to empirically test theoretical predictions against real-world data. Economic models provide the hypotheses, while statistical models offer the tools for empirical verification. This process is essential for validating theories, informing policy, and advancing our understanding of economic phenomena. This transformation is necessary because, while economic models can elucidate theoretical relationships, statistical models allow for the testing of these theories against real-world data. Several compelling reasons necessitate the conversion of economic models into their statistical counterparts:

1. **Quantification:** Economic models often provide qualitative predictions (e.g., an increase in price leads to a decrease in demand). Statistical models quantify these effects, offering precise estimates of the change in demand given a change in price.
2. **Empirical Validation:** Economic theories need testing. Statistical models bridge the gap between theory and reality, allowing us to empirically test the validity of a hypothesized economic relationship.
3. **Forecasting:** Statistical models built on sound economic principles facilitate prediction of future economic outcomes based on projected changes in the explanatory variables.
4. **Uncovering Hidden Factors:** The error term in the statistical model can reveal inadequacies of the original economic model and signal other potentially important factors that were not initially considered by the theory.

The transformation of an economic model into a statistical counterpart involves a few key steps:

1. **Specification:** Informed by the economic model, the first step is determining the precise functional form of the statistical model. This means deciding:
   - Which variables (from the economic model) to include.
   - Whether there are interactions between variables.
   - Any potential nonlinearities (e.g., using squared terms or logarithmic transformations)
2. **Data Collection:** The specified model dictates the relevant data needed. Economists must source reliable data on all the variables involved. The quality and availability of data significantly influence the reliability of the statistical model's results.
3. **Estimation:** This is where statistical techniques come into play. Using software tools, econometricians employ methods like Ordinary Least Squares (OLS) regression to estimate the unknown parameters (the β coefficients) of the model. These estimated coefficients provide insight into the magnitude and direction of the relationships between the variables.
4. **Hypothesis Testing:** Statistical models allow us to test our economic theories rigorously. Using tools like t-tests and F-tests, economists assess:
   - **Statistical Significance:** Are the estimated relationships statistically significant, or could they have occurred just by chance?
   - **Economic Significance:** Beyond statistical significance, we must consider the real-world impact of the magnitudes discovered. Does the model explain a meaningful portion of the variation in the data?
5. **Diagnostic Checks:** Before accepting the results, econometricians conduct diagnostic tests. This involves checking whether the model's assumptions hold, like the normality of the error term, absence of problematic multicollinearity among the variables, and more. If these assumptions are seriously violated, the model may need refinement.

Let's say, a researcher wants to empirically analyze the demand for coffee in a specific region. Using the economic model of demand as a guide, here's the process they might follow:

- **Specification:** The statistical model might be: $Q_d = \beta_0 + \beta_1 P_{coffee} + \beta_2 P_{avgIncome} + \beta_3 P_{tea} + \varepsilon$ (where P_{tea} is the price of tea, a possible substitute).
- **Data Collection:** Data would be gathered on coffee sales (Q_d), the price of coffee (P_{coffee}), average consumer income in the region ($P_{avgIncome}$), and tea prices (P_{tea}) over a period of time.
- **Estimation:** Using regression analysis, the researcher estimates the unknown coefficients (β).
- **Hypothesis Testing:** Were the estimated coefficients statistically significant? Did they align with the expectations of the economic model (e.g., negative coefficient for P_{coffee})?
- **Diagnostic Checks:** Did the model's errors seem randomly distributed? Was there any concerning correlation between the explanatory variables?

Consider another real-world application where policymakers are interested in understanding the impact of a soda tax on soda consumption. The economic model suggests that imposing a tax on soda would increase its price, leading to a decrease in quantity demanded.

- **Economic Theory:** Increase in soda tax leads to higher prices and thus lower quantity demanded.
- **Econometric Model:** ($Q_d = \beta_0 - \beta_1 Price_{soda} + \in$)
- **Statistical Analysis:** Data on soda prices (including the tax) and quantity sold are collected across various regions. Using an econometric estimation method, the effect of price on quantity demanded is estimated.
- **Findings and Policy Implications:** If the estimated coefficient for soda price is negative and statistically significant, it supports the hypothesis that the soda tax reduces consumption. This outcome can inform policymakers about the efficacy of the tax in achieving public health objectives.

From macroeconomics perspective, consider the question of how a tax on carbon emissions affects economic productivity. The economic model might posit that higher taxes on carbon lead to reduced emissions but might also impact productivity negatively. To empirically test this theory, we:

- Define productivity and carbon tax as our key variables, hypothesizing a negative relationship.

- Develop an econometric model that relates productivity to carbon tax and potentially other control variables (e.g., capital investment, labor force size).
- Incorporate randomness and other factors not explained by the model through an error term, specifying assumptions about this term's distribution.
- Use data on carbon taxes and productivity measures across different regions or times to estimate the model's parameters.
- Perform hypothesis testing to assess the statistical significance of the relationship between carbon taxes and productivity, thereby providing empirical evidence on the economic theory.

Another classic application of converting an economic model into a statistical model involves examining the impact of minimum wage increases on employment levels. The economic model might posit that higher minimum wages lead to lower employment due to increased labor costs. To test this theory, an econometric model could specify employment levels as a function of the minimum wage, controlling for other factors affecting employment. The statistical model would then include an error term to capture deviations from the model due to unobserved factors. By analyzing data on minimum wage changes and employment across different regions and time periods, researchers can estimate the model's parameters and test the economic theory against observed outcomes. While the transformation process may seem straightforward, it is crucial to acknowledge certain challenges and nuances:

- **Data Limitations:** The desired data may not be readily available or may be of poor quality, affecting the accuracy of the model.
- **Omitted Variable Bias:** The economic model may not account for all relevant factors. Omitting important variables in the statistical model can lead to biased results.
- **Structural Breaks:** Major economic events or significant policy changes can cause the underlying economic relationships to shift over time. Statistical models need to account for such breaks to maintain their validity.

Econometrics is as much an art as it is a science. Building reliable statistical models requires both a strong understanding of economic principles and the careful application of statistical methods. Success often relies on an iterative process of model refinement, data adjustments, and

careful exploration of alternative specifications until a model that is both statistically sound and economically meaningful is achieved.

2.1.3 Observations and the Error Term

Econometrics aims to test economic theories and models, often relying on observational data. This approach invariably introduces an element of uncertainty and variability unaccounted for within the theoretical structure of an economic model. Statistical models bridge this gap between economic theory and messy real-world data by incorporating this uncertainty via a key component: the error term.

A typical statistical model for regression analysis comprises several key components: the dependent variable, independent variables, coefficients, and the error term. Each component plays a critical role in the model's construction and interpretation:

- **Dependent Variable (Y):** This is the outcome or variable of interest that the model seeks to predict or explain. It is dependent on the independent variables. For example, in studying the impact of education on earnings, earnings would be the dependent variable.
- **Independent Variables (X):** These are the predictors or explanatory variables that are believed to have an effect on the dependent variable. In our example, the level of education, years of experience, and industry of employment could serve as independent variables.
- **Coefficients (β):** These values quantify the relationship between each independent variable and the dependent variable. A coefficient indicates the expected change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant.
- **Error Term (ε):** The error term represents the portion of the dependent variable that cannot be explained by the independent variables. It encompasses all other factors affecting the dependent variable that are not included in the model.

For example, the basic Keynesian consumption function posits:

$$C = a + bY$$

where:

C = Consumption expenditure

Y = Disposable income

a = Autonomous consumption (consumption when income is zero)

b = Marginal propensity to consume (MPC)

In order to test the validity of a theoretical model and estimate its parameters, we need to transform the economic model into a statistical one. Considering our consumption example, the statistical model would become:

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

Where β_0 and β_1 are population parameters (theoretical counterparts of a and b) to be estimated from the data, and ε is the error term.

The error term (ε) is a crucial component that encompasses all the complexities missing from our simplified economic model. It accounts for:

- **Measurement Errors:** Imprecision or inaccuracies in data collection of variables like consumption expenditure or income.
- **Omitted Variables:** Economic relationships are rarely isolated. There might be numerous minor influencing factors not explicitly included in the model.
- **Incorrect Functional Form:** The specified mathematical relationship (linear in this case) might be an approximation of a more complex, non-linear reality.
- **Stochastic Nature of Behavior:** Individual economic behavior can exhibit a degree of inherent randomness or unpredictability.

For our statistical analysis to produce reliable results, we make certain assumptions about the error term:

- **Zero Mean:** On average, the errors will balance out across observations ($E[\varepsilon] = 0$). This implies no systematic over or underestimation of the dependent variable.
- **Constant Variance (Homoscedasticity):** The error term has the same variance ($Var[\varepsilon] = \sigma^2$) across across all observations.
- **No Autocorrelation:** The error terms are uncorrelated with each other; errors in one period have no bearing on errors in other periods ($cov(\varepsilon_i, \varepsilon_j) = 0$ *for* $i \neq j$).
- **Normality:** The error term follows a normal distribution. While not strictly necessary, this assumption helps with hypothesis testing and confidence intervals.

A well-constructed statistical model allows us to isolate the effect of specific economic variables while accounting for uncertainty and other unobserved factors. To illustrate the application of regression analysis and the role of the error term, consider two examples from economics:

As first example, consider the impact of education on earnings. The economic model, based on the human capital theory suggests that an individual's earnings increase with higher levels of education, *ceteris paribus*. To test this theory, a regression model could be specified as:

$$\text{Earnings} = \beta_0 + \beta_1 \text{ (Years of Education)} + \epsilon$$

Here, the dependent variable is earnings, and the independent variable is years of education. The error term represents factors other than education that might influence earnings, such as experience, innate ability, and economic conditions.

For second example, consider the theory of wage determination. It suggests that current wage is determined majorly by education level, experience, skills, and other factors. The corresponding statistical model for predicting potential future wage would be:

$$\text{Wage} = \beta_0 + \beta_1 \text{(Education)} + \beta_2 \text{ (Experience)} + \beta_3 \text{ (Skills)} + \varepsilon$$

The error would encapsulate factors like unobserved ability/ motivation, discrimination, regional wage differences, measurement errors in experience, etc.

In econometrics, the construction of a statistical model to examine an economic model is a meticulous process that requires careful consideration of the dependent and independent variables, the estimation of coefficients, and, crucially, the understanding of the error term. The error term is not merely a catch-all for unexplained variation but a critical component that influences the model's specification, estimation, and inference.

2.1.4 Estimating the Parameters of the Econometric Model

The estimation of parameters in an econometric model is a fundamental aspect of regression analysis. These models comprise parameters—constants that shape the model's specific form and predictions. Estimating these parameters from available data is crucial for the model to provide accurate, reliable predictions and inferences about economic phenomena. For example, take a model examining the relationship between household and expenditure:

$$\text{Expenditure} = \beta_0 + \beta_1 (\text{Income}) + \varepsilon$$

Here, expenditure is the dependent variable, income is the independent variable (also called the regressor or explanatory variable), and ε is the error term that encompasses all factors influencing expenditure that are not explicitly included in the model. The parameters of this model are:

- β_0 (intercept): the expected expenditure when income is zero.
- β_1 (slope): the expected change in expenditure for a one-unit change in income.

The central goal of parameter estimation is to find the *best* values of β_0 and β_1 that describe the observed economic data. Best, in this context, depends on the estimation method selected. Let's say that we collect a sample of household income and expenditure data. The estimation aims to find a line (defined by its intercept and slope) that most closely reflects the underlying relationship observed in the data. Of course, the line is unlikely to pass through every data point perfectly, representing the role of other factors and potential measurement errors captured within the error term.

Estimating the parameters of an econometric model involves using statistical methods to infer the values of these parameters that best explain the observed data. The accuracy and reliability of econometric analyses hinge on the precision of these estimations, as they directly influence the model's predictive power and the validity of the inferences drawn from it. Several methods exist for the estimation of parameters in econometric models. The most prevalent techniques include:

Ordinary Least Squares (OLS): OLS remains the workhorse of econometric estimation. Its principle is to find the parameter values that minimize the sum of squared errors (residuals), i.e., the vertical distances between the actual data points and the line defined by the estimated parameters. OLS is computationally straightforward and has desirable statistical properties under specific assumptions. It makes the following assumptions:

- Linearity: The relationship between the dependent and independent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of the error terms is constant across observations.

- Normality: The error terms are normally distributed (especially relevant for hypothesis testing).

These assumptions also mark the limitations of the technique:

- Non-linearity, heteroscedasticity, or autocorrelation in the data can lead to biased and inefficient OLS estimates.
- OLS is sensitive to outliers and can be influenced by multicollinearity.

OLS is apt for situations where the relationship between variables is linear and the data meet the assumptions of independence, homoscedasticity, and normality. It is extensively used in economics for demand analysis, forecasting, and policy evaluation. For example, economists analyze the price elasticity of demand for a product using a model where quantity demanded is the dependent variable and price is the independent variable. OLS estimation helps quantify how much demand changes in response to price changes. Also, in analyzing the factors affecting housing prices, an economist might use OLS to estimate a linear regression model where the dependent variable is the price of houses and the independent variables include house size, location, age, and other relevant factors. This analysis can inform policy decisions regarding housing affordability and urban planning.

Maximum Likelihood Estimation (MLE): MLE involves finding the parameter values that maximize the probability (likelihood) of observing the given data, conditional on a specific probability distribution for the errors. MLE is flexible in accommodating different distributional assumptions and is generally statistically efficient. The method assumes:

- A specified probability distribution for the error terms.
- Independence of observations.
- Adequately large sample size for the Law of Large Numbers to hold.

The limitations of this technique are:

- MLE requires the specification of a probability distribution for the error terms, which might be incorrect.
- It can be computationally intensive for complex models.
- MLE estimates can be biased in small samples.

MLE is particularly useful for non-linear models and models where the error terms follow a non-normal distribution. It is widely used in duration analysis, count data models, and for estimating the parameters of discrete choice models. An example of MLE application is in the analysis of unemployment duration. By assuming that the duration of unemployment periods follows a particular probability distribution (e.g., exponential or Weibull), economists can use MLE to estimate the parameters of this distribution, thereby gaining insights into the dynamics of job search behavior and the effectiveness of labor market policies. MLE estimation is also employed in asset pricing models (e.g., the Capital Asset Pricing Model), where the expected return of an asset is a function of its systematic risk. OLS and MLE often find application in such models.

Instrumental Variables (IV): IV estimation is employed when an independent variable is correlated with the error term (a problem known as endogeneity). IV techniques use "instruments," variables correlated with the independent variable but uncorrelated with the error term, to obtain consistent estimates.

Generalized Method of Moments (GMM): GMM is a powerful and versatile framework that encompasses many estimation techniques as special cases. It utilizes "moment conditions" (expectations involving the data and model parameters) to derive parameter estimates, making it more robust to assumptions than OLS and MLE. GMM is a flexible estimation technique that generalizes the method of moments by allowing for the use of multiple moment conditions. It is particularly useful when the model's assumptions do not perfectly match the characteristics of the data. The method only assumes that:

- The model's specified moment conditions are valid.
- The instruments used are valid (uncorrelated with the error term).

It suffers from the following limitations:

- The choice of instruments and moment conditions can significantly affect the estimates.
- GMM estimates can be inefficient if the wrong moment conditions or instruments are used.

GMM is applicable in situations where traditional assumptions (such as normality or homoscedasticity) do not hold, or when dealing with simultaneous equations models. It has been employed in financial

economics to estimate risk-return tradeoffs and in macroeconomics to estimate dynamic stochastic general equilibrium (DSGE) models.

For instance, GMM is used in financial economics to estimate the parameters of asset pricing models. By using historical data on asset returns and applying relevant moment conditions, researchers can assess the validity of different asset pricing theories and their implications for investors' portfolio choices. In macroeconomics, researchers investigate the determinants of economic growth by constructing models where GDP growth is the dependent variable, and explanatory variables may include investment rates, education levels, and technological progress. IV and GMM techniques might be needed to address endogeneity among these variables. Deciding which estimation method to use requires careful consideration of several factors:

- **Distributional Assumptions:** MLE is often tied to a specific distributional assumption about the error term (e.g., normal distribution). OLS, although less sensitive to the error distribution, still performs best with normally distributed errors.
- **Endogeneity:** Problems like omitted variables, measurement errors, or simultaneity cause endogeneity, invalidating OLS. IV techniques address such endogeneity issues.
- **Computational Complexity:** MLE and GMM can be computationally more demanding than OLS, especially with complex models.

It is vital to be mindful of the potential limitations of econometric estimation:

- **Model Misspecification:** If our underlying theoretical model is flawed (e.g., excluding relevant variables), even a correctly estimated model leads to biased results.
- **Data Quality:** Poor quality data with measurement errors or outliers can severely distort estimates.
- **Structural Change:** If the underlying relationships change over time, using the entire dataset for estimation can provide misleading results.
- **Interpretation:** It is crucial to interpret estimated parameters within the context of the economic theory and limitations of the model.

The judicious choice of estimation techniques, coupled with a careful interpretation of the estimated parameters, allows us to extract valuable insights from economic data to inform better decision-making in both policy and business realms. While each estimation method has its own set of assumptions and limitations, the choice of method depends on the model specification, the nature of the data, and the research objectives. Understanding these techniques in depth, along with their applicability and limitations, is essential for conducting rigorous econometric analysis and for the critical evaluation of econometric research.

2.1.5 The Ordinary Least Squares Estimation Method

Ordinary Least Squares (OLS) seeks to find the best-fit line through a set of data points. This line represents a linear model that describes the relationship between an outcome variable (dependent variable) and one or more predictor variables (independent variables). Imagine a scatterplot depicting the relationship between an individual's income (outcome variable) and their education level (predictor variable). The OLS method attempts to draw the "best" line summarizing the pattern in this data.

The OLS method works by minimizing the sum of squared errors, also known as residuals. Residuals are the vertical distances between the actual data points and the fitted regression line. In other words, it minimizes the sum of the squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation. The aim is to find the line that results in the smallest possible sum of these squared errors.

The use of OLS as an econometric model to estimate (or predict) the values of an outcome variable must be preceded by analyzing the relationship between the outcome and the predictor variables from within the theory and then, by examining the relationship between the two by plotting the data points from the sample collected. Only if the theory postulates a linear relationship between the two variables or, in absence of a clear formulation, if the scatter plot exhibits a pattern that can be reasonably approximated by a straight line, can the OLS method be used in any statistically meaningful way.

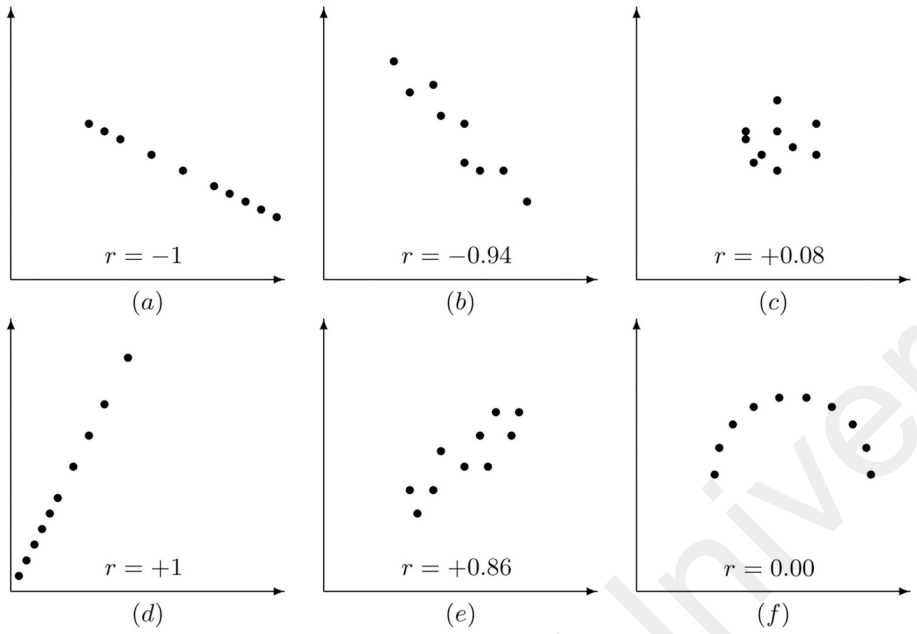


Figure 1: Scatterplots and Correlations (This Photo by Unknown Author is licensed under CC BY-SA)

Consider an example from the macroeconomic theory. The Keynesian consumption function says that consumption is dependent primarily on the level of income. So mathematically the relationship between the two can be simply put as:

$$c = f(y)$$

Where, c is the consumption expenditure and y is the level of income. But the theory does not specify any specific functional form. This can be examined only by collecting a sample of consumption expenditures and income levels of households. The sample datapoints are then scatter plotted in order to infer a pattern among them. If the scatter points are within a tight enough narrow band, the relationship between consumption and income can be said to be approximated by a straight line–the sample consumption function:

$$c = b_o + b_1 y$$

This sample consumption function is the representative of the true population consumption function, given by:

$$C = \beta_0 + \beta_1 Y$$

But the mathematical form of the consumption function may not capture all factors that influence the consumption decision of the household–even though the income may well be the single most important

factor in that decision. To account for the uncertainty, reflected by the spread of datapoints in the scatterplot, an error term is added to the equations converting them into statistical models:

$$Population: C = \beta_0 + \beta_1 Y + u$$

$$Sample: c = b_o + b_1 y + e$$

For econometric estimation, a sample will be collected and the estimation equation will take the form:

$$c_i = b_o + b_1 y_i + e_i$$

Where the subscript, *i*, represents the values observed for the i^{th} household. Thus, the error in estimation can be represented as:

$$e_i = c_i - (b_o + b_1 y_i)$$

If the sample collected is of the size *n*, then the OLS estimators of β_0 and β_1 are calculated by minimizing the sum of squared errors:

$$min: e_i^2 = \sum_{i=1}^{n} \left(c_i - \hat{c}_i \right)^2$$

Where ($\hat{c}_i$)is the predicted value of c_i , given by $\hat{c}_i = b_0 + b_1 y_i$. The minimization leads to the OLS normal equations, which can be solved to obtain the values of the parameters. We will show how to derive the expressions for the parameters of a simple linear regression equation using the generalized form:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

The sum of squared errors would be:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i \beta_1 \right)^2$$

$$= \sum_{i=1}^{N} \left(y_i^2 + \beta_0^2 + x_i^2 \beta_1^2 - 2 x_i y_i \beta_1 + 2 x_i \beta_0 \beta_1 \right)$$

$$= \sum y_i^2 + \beta_0^2 N + \beta_1^2 \sum x_i^2 - 2\beta_0 \sum y_i - 2\beta_1 \sum x_i y_i + 2x\beta_0 \beta_1 \sum x_i$$

To find the minima, we take out first order partial derivatives of the sum of squared errors:

$$\frac{\partial S}{\partial \beta_0} = 2N\beta_0 - 2\sum y_i + 2\sum x_i \beta_1$$

$$\frac{\partial S}{\partial \beta_1} = 2\sum x_i^2 \beta_1 - 2\sum x_i y_i + 2\sum x_i \beta_0$$

and set them to zero in order to find out the points (b_0, b_1) that minimize the function:

$$2\left(\sum y_i - Nb_0 - \sum x_i b_1\right) = 0$$

$$2\left(\sum x_i y_i - \sum x_i b_0 - \sum x_i^2 b_1\right) = 0$$

Where b_0 and b_1 are the estimates of β_0 and β_1, respectively. Rearranging the equations will yield:

$$Nb_0 + \sum x_i b_1 = \sum y_i$$

$$\sum x_i b_0 + \sum x_i^2 b_1 = \sum x_i y_i$$

These normal equation can be solved as a system of simultaneous equations in two variables, b_0 and b_1.We multiply both sides of the first equation by Σxi and both sides of the second equation with N to get:

$$N\sum x_i b_0 + \left(\sum x_i\right)^2 b_1 = \sum x_i \sum y_i$$

$$N\sum x_i b_0 + N\sum x_i^2 b_1 = N\sum x_i y_i$$

Subtracting the first equation above from the second yields:

$$\left[N\sum x_i^2 - \left(\sum x_i\right)^2\right] b_1 = N\sum x_i y_i - \sum x_i \sum y_i$$

Solving for b_1 yields:

$$b_1 = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2}$$

Which can be alternatively expressed as:

$$b_1 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}$$

Also, dividing both sides of the first normal equation by N yields:

$$b_0 = \overline{y} - b_1 \overline{x}$$

Consider a simple dataset to illustrate OLS estimation and interpretation. We wish to estimate the relationship between education and income, $Y = \beta_0 + \beta_1 X + u_i$, for which we collect data from ten individuals about their monthly income (Y) and years of education (X). Table below lists the data points.

Education (years) [X]	Income ('000s) [Y]
0	0
2	5
4	10
6	15
8	20
10	25
12	30
14	35
16	40
18	45
20	50

The sample linear regression equation would be: $y_i = b_0 + b_1 x_i + e_i$. The sample means are: $\bar{x} = 10$; $\bar{y} = 22.45$. Putting these values in the OLS normal equations, gives us: $b_0 \cong 0$; $b_1 = 2.25$. This means that, on average, each additional year of education is associated with an increase of Rs.2250 in income. The intercept ($b_0 \cong 0$) can be interpreted as the expected income (in this case, Rs.0 per month) for an individual with zero years of education. This estimate quantifies the marginal effect of education on income, suggesting a positive and significant relationship between these variables in our model.

Let us plot the regression line along with the data points to visually assess the fit of our model. The graph will show the observed data points (income as a function of education) and the estimated regression line based on the calculated OLS coefficients.

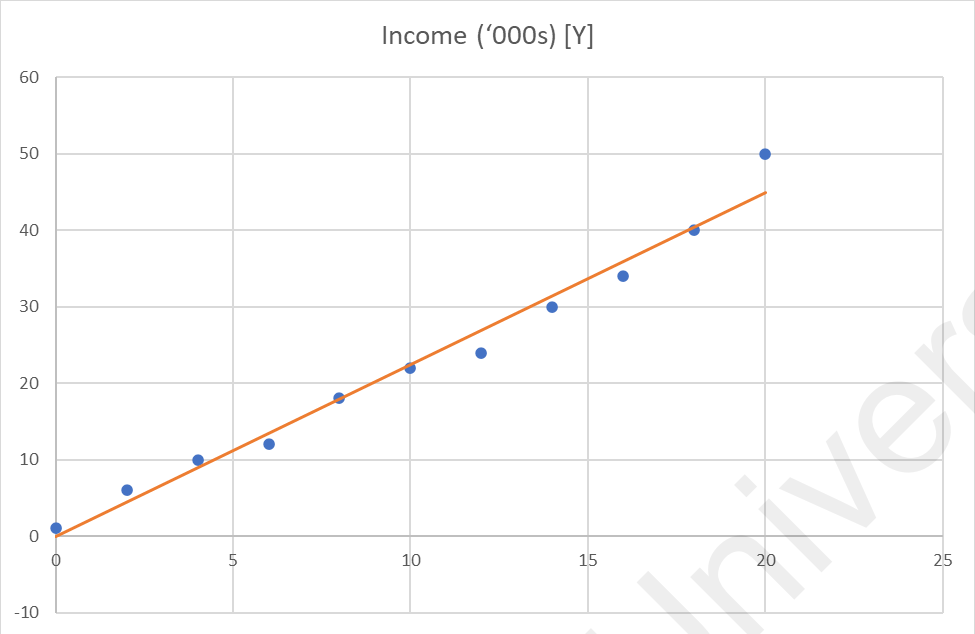


Figure 2: The Fitted Regression Line

The plot illustrates the observed data points (in blue) representing the relationship between income and years of education. The OLS regression line (in orange) provides a linear approximation of this relationship, based on the calculated estimators. This visual representation confirms our earlier interpretation: as education increases, income also tends to increase, with the rate of increase being approximately Rs.2250 for each additional year of education.

The OLS estimators have several important properties. They are the best, unbiased estimators among the class of linear estimators–a property commonly summarized as BLUE (best, linear, unbiased estimator):

- Efficiency (Best): Among all linear and unbiased estimators, OLS estimators have the smallest variance.
- Linearity: They are linear functions of the dependent variable. Alternatively, they can be interpreted as being linear in parameters.
- Unbiasedness: On average, the estimators will equal the true parameter values.
- Consistency: As the sample size increases, the estimators converge in probability towards the true parameter values.

Recent research has shown that if the underlying data is normally distributed, the OLS estimators are the best unbiased estimators among all possible estimators. But these properties of the OLS estimators are

predicated on a set of assumptions. For OLS to produce reliable results, certain assumptions must be met. The classical OLS assumptions are:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. No Perfect Multicollinearity: The independent variables are not perfectly linearly related.
3. Zero Conditional Mean: The expected value of the error term, given the independent variable, is zero.
4. Homoscedasticity: The error terms have constant variance.
5. No Autocorrelation: The error terms are not correlated with each other.
6. Normality of Errors (optional for inference): The error terms are normally distributed.

While OLS is widely used, it has limitations: (i) applies only to linear models, (ii) is sensitive to outliers (which can significantly affect the estimates), and (iii) violations of its assumptions can lead to biased, inconsistent, or inefficient estimators.

After calculating the OLS estimates, it is essential to conduct diagnostic tests to check for violations of OLS assumptions (e.g., residual plots to assess normality, homoscedasticity, etc.). Also, OLS estimates come with standard errors. These allow for hypothesis testing and construction of confidence intervals to determine the statistical significance of the relationship.

OLS is a fundamental tool in econometrics, offering a simple yet powerful method for linear regression analysis. Understanding its assumptions, limitations, and proper application is crucial for accurate modeling and interpretation of relationships between variables.

2.1.6 The Maximum Likelihood Estimation Method

Maximum Likelihood Estimation is a statistical technique used to estimate the parameters of a model by maximizing the likelihood function, which represents the joint probability distribution of the observed data given the parameters. The fundamental principle behind MLE is to find the parameter values that make the observed data most probable.

The OLS method estimates parameters by minimizing the sum of the squared residuals (differences between observed and predicted values). In contrast, MLE estimates parameters by maximizing the likelihood function, which is the probability of observing the sample data given the parameters. The key difference lies in the optimization criterion: OLS minimizes residuals, while MLE maximizes likelihood.

While the OLS relies heavily on assumptions related to the linear regression model, such as linearity, homoscedasticity (constant variance), and normality of error terms. MLE is more flexible in terms of the distribution of the error terms, allowing for models where the error terms are non-normal or heteroscedastic, or cases when the dependent variable is binary or count data. MLE provides consistent, asymptotically efficient, and normally distributed estimates under weaker assumptions than OLS, making it more robust and flexible. MLE only assumes that the specified model is the correct model with constant and finite parameters, and that the underlying the data are independent and identically distributed (i.i.d.).

Following are the steps for estimating the parameter values of the specified model using the Maximum Likelihood Estimation method:

- Specify the probability distribution of the data.
- Write down the likelihood function, $L(\theta|y)$, and the log-likelihood function, $lnL(\theta|y)$.
- Take the derivative of the log-likelihood function with respect to the parameters, θ.
- Set the derivative to zero and solve for θ to find the MLE estimates, $\hat{\theta}$.

The likelihood function, $L(\theta|y)$, is the joint probability distribution of the observed data, y, given the parameters, θ. The log-likelihood function, $lnL(\theta|y)$, is often used instead for computational convenience. The MLE estimates, $\hat{\theta}$, are the values of θ that maximize the log-likelihood function. Consider a simple linear regression model:

$$y = \beta_0 + \beta_1 x + \varepsilon, \ where \varepsilon \sim N(0, \sigma 2)$$

Given a sample of data $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, the likelihood function is:

$$L(\beta_0, \beta_1, \sigma^2 \mid y) = \prod \left[\frac{1}{\sqrt{2\pi\sigma^2}} \right] exp \left[-\frac{(y_i - \beta_0 - \beta_1 x_i) 2}{2\sigma^2} \right]$$

Taking the logarithm, we get the log-likelihood function:

$$\ln L\left(\beta_0, \beta_1, \sigma^2 | y\right) = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \Sigma \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

Taking the derivative with respect to β_0, β_1, and σ^2, setting them to zero, and solving for β_0, β_1, and σ^2, we get the MLE estimates. Once the parameters are estimated, their statistical significance and the goodness-of-fit of the model can be evaluated through tests like the likelihood ratio test.

Consider an example: suppose that the marks scored by students in an introductory course in econometrics are distributed normally. A random sample of ten students had the following scores (out of a maximum of 200):

Student	1	2	3	4	5	6	7	8	9	10
Score (X)	113	127	159	138	180	117	131	149	152	171

We need to identify the maximum likelihood function in order to obtain the maximum likelihood estimate of the mean score. Since the scores are distributed normally, the probability density function of the mean score will be:

$$f\left(x_i; \mu, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} exp \left[-\frac{\left(x_i - \mu\right)^2}{2\sigma^2} \right] for -\infty < x, \mu < +\infty \, and \, 0 < \sigma < \infty$$

Therefore, the likelihood function will be:

$$L\left(\mu, \sigma\right) = \sigma^{-n} \left(2\pi\right)^{-\frac{n}{2}} exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(x_i - \mu\right)^2 \right]$$

$$for -\infty < \mu < +\infty \, and \, 0 < \sigma < \infty$$

Using the maximizing procedure outlined earlier, we get the MLE estimator of μ as:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

Thus, the sample yields the maximum likelihood estimate of the mean score as:

$$\hat{\mu} = \frac{1}{10}(113 + 127 + ... + 171) = 144.5$$

The Maximum Likelihood Estimation method is a powerful and flexible tool in econometrics, providing robust and efficient estimates under weaker assumptions than the Ordinary Least Squares method. However, MLE has its own set of limitations: it may not provide accurate estimates for small samples, it may be computationally intensive, especially for complex models, and it may be biased if the model is mis-specified. Also, like other estimation methods, MLE can be prone to overfitting, especially in models with a large number of parameters relative to the sample size. Applying MLE requires careful specification of the probability distribution and is, generally, computationally intensive.

2.1.7 Other Statistical-Econometric Models

While linear regression is the workhorse of statistical analyses, there are several other advanced modeling techniques with general and specialized application:

1. **Generalized Linear Models (GLMs):** GLMs extend linear regression to handle dependent variables that don't neatly follow a normal distribution. They do this through using a *link function* that connects the linear part of the model to the mean of the dependent variable's distribution and specifying the probability distribution (e.g., binomial, Poisson, gamma). Examples include:
   a. Logistic Regression: For binary outcomes (e.g., employed vs. unemployed, purchase vs. no purchase). Uses a logit link function and a binomial distribution.
   b. **Poisson Regression:** For count data (e.g., the number of accidents in a month, number of patents a firm files). Employs a log link function and the Poisson distribution.
2. **Time Series Models:** Specialize in analyzing data with a temporal component, where the order of observations matter. Account for trends, seasonality, and potential autocorrelation (correlation between a variable and its past values). Examples are:

a. **ARIMA (Autoregressive Integrated Moving Average):** Models a variable as a function of its own past values and past errors.

b. **ARCH/GARCH (Autoregressive Conditional Heteroskedasticity):** Focus on modeling changing variance (volatility) over time, often seen in financial data.

3. **Limited Dependent Variable Models:** Handle situations where the dependent variable is not continuous across all values, e.g., it might be binary, censored, or truncated. Examples are:

   a. **Probit Regression:** Similar to logistic regression but uses the cumulative standard normal distribution function as its link.

   b. **Tobit Regression:** For data where a portion of observations clump at a certain value (frequently zero). For instance, modeling charitable donations (many will be zero, some will be positive amounts).

4. **Discrete Choice Models:** Model decisions individuals or firms make among a set of alternatives. Based on utility maximization concepts. Examples include:

   a. **Multinomial Logit:** Used when choices are unordered (e.g., choice of transportation mode: car, bus, train).

   b. **Conditional Logit:** Useful when choices can be grouped (e.g., picking a specific brand within a product category)

5. **Systems of Equations:** Allow modeling several relationships simultaneously, where the dependent variable in one equation might be an independent variable in another. Accounts for interdependence between outcomes. Examples are:

   a. **Seemingly Unrelated Regressions (SUR):** A system of seemingly separate regression equations where the error terms might be correlated across equations.

   b. **Simultaneous Equation Models:** Explicitly models structural feedback loops, where variables can influence each other mutually.

While OLS and MLE are cornerstones of regression analysis, several other estimation methods deserve consideration:

1. **Method of Moments (MM):** Equates theoretical moments (e.g., mean, variance) of the population with their sample counterparts. Parameters are estimated by solving the resulting system of equations. **Example:** consider a simple model $Y = \beta X + \varepsilon$. The first

population moment is $E[Y] = \beta E[X]$. Its sample equivalent is $\bar{Y} = \beta \bar{X}$. Solving this equation gives an estimate for β.

a. **Pros:** Can be computationally simpler than MLE.

b. **Cons:** Often less efficient (higher variance) than MLE. May not be feasible if the number of moments exceeds the number of unknown parameters.

2. **Bayesian Estimation:** A fundamentally different philosophical approach. It treats parameters as random variables with prior distributions (beliefs about the parameter before seeing the data). The posterior distribution (updated belief after incorporating data) is derived using Bayes' Theorem. It involves specifying the prior distribution, the likelihood function, and calculating the posterior distribution. Estimates are often based on features of this posterior, like the mean.

   a. **Pros:** Incorporates prior knowledge in a natural way. Provides a full probabilistic characterization of the parameters.

   b. **Cons:** Computationally intensive, particularly with complex models. Results can be sensitive to prior distribution choice.

3. **Robust Regression:** Methods designed to be less sensitive to outliers or deviations from the assumptions of traditional regression models. It employs several techniques, such as:

   a. **M-Estimators:** Minimize a function that down-weights the influence of outliers.

   b. **Least Absolute Deviations (LAD):** Minimizes the sum of absolute residuals rather than squared residuals.

The choice of method employed depends on the specific dataset, research questions, and the theoretical underpinnings of the model.

2.1.8 Summary

Economic models provide a theoretical framework, while statistical models are tools for empirical testing of these theories using data. Econometric models include an error term to account for the randomness and other factors not captured by the model.

Regression analysis is a statistical method used to explore the relationship between a dependent variable and one or more independent variables. It is versatile and can be applied to various economic questions, like the impact of education on earnings or factors influencing consumer spending.

Simple Linear Regression is the basic form of regression that assumes a linear relationship between the dependent and independent variables. It's useful for understanding the impact of changes in one variable on another. Multiple Regression Analysis is an extension of simple linear regression that includes several independent variables to provide a more complex and nuanced understanding of economic dynamics.

Parameters are estimated using various methods with each method having its own assumptions and limitations. Ordinary Least Squares (OLS) is a method that estimates the parameters of a regression model by minimizing the sum of squared residuals. Maximum Likelihood Estimation (MLE) method estimates parameters by maximizing the likelihood of observing the sample data, given the parameters.

Beyond the linear model, there are logistic regressions for binary outcomes, time series models for data over time, and more complex models like ARIMA and GARCH for specific types of dependency in data.

Issues such as multicollinearity, heteroscedasticity, and autocorrelation can affect the reliability of regression models. Regression analysis requires clean data, appropriate models, and robust statistical techniques.

2.1.9 Keywords

Regression Analysis: A statistical approach for examining the relationships between variables, used to model and predict the association between a dependent variable and one or more independent variables.

Simple Linear Regression: Assumes a linear relationship between the dependent and independent variables. The linear regression equation is given by $y = c + mx$, which helps quantify the impact of changes in the independent variable on the dependent variable.

Multiple Regression Analysis: Extends simple linear regression by incorporating multiple independent variables to explain the dependent variable, allowing for a more detailed exploration of complex economic relationships.

Ordinary Least Squares (OLS): A method that estimates regression model parameters by minimizing the sum of squared differences between the observed and predicted values.

Maximum Likelihood Estimation (MLE): Estimates parameters by finding the values that maximize the likelihood of the observed data.

Economic vs. Statistical Model: Economic models offer theoretical frameworks, while statistical models apply these theories to empirical data to test relationships.

Error Term: Represents unexplained variance in the dependent variable not accounted for by the independent variables in the model.

Estimating Parameters: Involves selecting the best method based on the model's characteristics, such as OLS or MLE, and then using the chosen method to derive estimates.

Model Specification and Diagnostics: Critical for ensuring the reliability of regression models. This includes proper data preparation, exploratory data analysis, model fitting, and diagnostic checks to address any violations of assumptions.

2.1.10 Self-assessment Questions

1. What is regression analysis and what is its primary purpose in econometrics?
2. Define a dependent variable and give an example in the context of regression analysis.
3. What are independent variables and how do they function in a regression model?
4. What does the term linear relationship imply in the context of simple linear regression?
5. Explain the equation $y = c + mx$ in the simplest terms, identifying what each component represents.
6. How does multiple regression analysis differ from simple linear regression?
7. Describe the Ordinary Least Squares (OLS) estimation method and its objective.
8. What is the Maximum Likelihood Estimation (MLE) method and when is it used?
9. Contrast economic models with statistical models in the context of econometrics.
10. What is the *error term* in a regression model and what does it account for?

11. Why is it important to ensure that the assumptions of regression analysis are met?
12. List the key assumptions that must hold true for linear regression to produce reliable results.
13. What steps would you take to prepare data for regression analysis?
14. What are some of the diagnostic checks that should be performed after fitting a regression model?
15. How can regression analysis be applied in the real world? Provide at least two examples.

2.1.11 References

1. **Introduction to Econometrics** by James H. Stock and Mark W. Watson: This book is well-suited for those who have a grasp of the conceptual part of economics and are ready to delve into econometrics. It's recommended to have a background in statistics before tackling this book, as it can be quite dense for complete beginners.
2. **Econometric Analysis of Cross Section and Panel Data** by Jeffrey M Wooldridge: This book is better suited for advanced students, such as those in a Ph.D. program. It's an excellent reference for understanding cross-section and panel data methods in detail.
3. **Microeconometrics Using Stata** by A. Colin Cameron and Pravin K. Trivedi: Ideal for those looking to bridge the gap between econometric theory and Stata application, this book provides detailed instructions on Stata commands within the context of econometric analysis.
4. **Basic Econometrics** by Damodar Gujarati: Known for its simple approach and explanation of econometrics without heavy mathematical focus. This book is beneficial for those taking economics courses and aims to blend current research with traditional econometric theory.
5. **Econometric Analysis** by William H. Greene: This is one of the most prominent textbooks for graduate-level econometrics and offers a hands-on approach to learning fundamental econometric techniques and models, making it an essential reference for researchers and practitioners.

Lesson 2.2 – Multiple Linear Regression

Structure

2.2.1 Introduction to Multiple Regression Analysis

Multiple regression analysis is a statistical technique used to analyze the relationship between multiple independent (predictor or explanatory) variables and a single dependent (response or outcome) variable. It extends simple linear regression, which analyzes the relationship between two variables, to a more complex scenario where the effect of multiple factors on a particular outcome is examined.

By including multiple independent variables, we can create a more comprehensive and realistic model to make predictions or understand the relationships between variables. This technique is crucial in economics, as well as in various other fields like finance, healthcare, marketing, and social sciences, where understanding multifactorial influences is essential. In economics and beyond, multiple regression analysis is used in numerous contexts:

- **Economics**
  - Predicting house prices (dependent variable) based on various factors such as the size of the house, location, number of bedrooms, and age of the house (independent variables).
  - Estimating the impact of various factors like interest rates, consumer confidence, and unemployment rates on economic indicators such as GDP growth.

- Investigating how factors like education level, experience, gender, location, industry, etc., collectively influence an individual's wage.
- Modeling how income, prices, interest rates, and household demographics might affect consumer spending.
- Determining the impact of company financials, macroeconomic indicators, and market sentiment on a stock's price.
- Predicting a product's demand based on its price, advertising expenditure, competitor prices, and seasonal factors.

- **Marketing**
  - Determining the impact of advertising spending on different platforms (TV, radio, social media) on total sales (dependent variable).
  - Evaluating how various marketing activities (advertising spend across different media, pricing strategies, etc.) influence sales or brand awareness.
- **Education**
  - Analyzing the factors influencing students' test scores (dependent variable), such as hours of study, attendance, and parental involvement (independent variables).
- **Finance**
  - Analyzing the influence of different financial ratios and market conditions on stock prices or company valuation.
- **Medicine and Healthcare**
  - Understanding how different demographic, lifestyle, and medical variables affect health outcomes.
  - Determining the influence of age, lifestyle, and genetics on disease risk.
- **Sociology**
  - Studying how various factors affect crime rates, educational outcomes, or voting behaviors.
- **Environmental Science**
  - Modeling pollution levels as a function of industrial activity, population density, and weather patterns.

Multiple regression makes the same set of assumptions as the simple linear regression but comes with an added complication of multicollinearity-potentially high correlation among independent variables in the model. Let's take an example to understand the formulation of a multiple regression model—how education, experience, and gender influence wages. In this case:

Dependent Variable (Y): Wage

Independent Variables (X's):

X_1 = Years of Education

X_2 = Years of Work Experience

X_3 = Gender (represented as dummy variable)

Then this study model can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Where:

β_0: The intercept (the expected value of Y when all independent variables are zero)

β_1, β_2, β_3: The slope coefficients; represent the expected change in Y for a one-unit increase in the corresponding independent variable, holding all other independent variables constant.

ϵ: The error term (unobserved factors that affect Y)

This form can be extended to include any number of independent or predictor variables as needed in the study:

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

The standard form of a multiple regression model is essentially the equation given above. It succinctly expresses the linear relationship between the dependent variable and multiple independent variables. Since a typical dataset for such studies involves a large number of cases (e.g., number of respondents in a survey), writing the model equation for each case becomes unwieldly very rapidly. For example, the proposed study outlined above, a researcher will typically collect data from hundreds (even thousands) of individuals. The model will thus look like:

$$Y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ij} + \beta_3 X_{ij} + \epsilon_i$$

Where the subscripts (*i, j*) represent data of i^{th} variable collected from j^{th} individual. Expanded, for *n* number of respondents, as:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} + \grave{o}_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} + \grave{o}_2$$

.

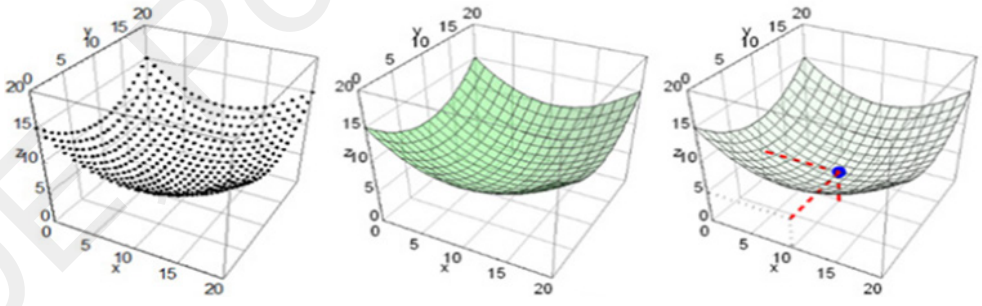
.

.

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \beta_3 X_{3n} + \grave{o}_n$$

Thus, X_{22} would mean years of experience data for the second respondent. It is easy to see that writing hundreds of such datalines is both cumbersome and unweildly. It also complicates calculations for estimating the values of the parameters. Situation componds when the number of independent variables in the model is also large. In such cases, it is much more convenient to use the matrix notations for expressing the regression equation. Use of matrices not only compresses data representation but also eases parameter estimation.

This can be visualized as a multidimensional space where each dimension represents one of the variables in the model. The multiple regression model would be represented by a hyperplane (a line in two dimensions, a plane in three dimensions, and so on) that best fits the data points in this multidimensional space. The coefficients (**β** values) determine the orientation and position of this hyperplane, optimizing the prediction of the outcome variable (***Y***) based on the predictor (***X***) variables in the model.



The goal of multiple regression is to estimate the parameters (**β**s) and examine their (i) significance (are the independent variables statistically significant in explaining the dependent variable?), direction (is the relationship positive or negative?), and magnitude (how big of an effect does each independent variable have on the dependent variable?).

2.2.2 Estimation of Model Parameters in Multiple Regression

We will use matrix algebra to estimate the model parameters of a simple linear regression relationship with two variables:

$$y_1 = \beta_0 + x_1\beta_1 + e_1$$
$$y_2 = \beta_0 + x_2\beta_1 + e_2$$
$$\vdots$$
$$y_n = \beta_0 + x_n\beta_1 + e_n$$

Since the result will be in the form of matrix equations, it can be easily generalized for use in multiple regression models. The model above can be represented in matrix algebra form as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \beta b_0 + \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \beta_1 + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

Or

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

We can write the expressions above in a more compact form by constructing vectors of *N*-dimensions:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ 1 \end{bmatrix} \boldsymbol{x}_2 = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ \cdot \\ e_n \end{bmatrix}$$

And, thus, rewrite this as:

$$\mathbf{y} = \mathbf{x}_1\boldsymbol{\beta}_0 + \mathbf{x}_2\boldsymbol{\beta}_1 + \mathbf{e}$$

This can be further compacted if we create:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{pmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix} \quad and \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Using this notation, we can rewrite as:

$$\boldsymbol{x}_1\beta_0 + \boldsymbol{x}_2\beta_1 = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \boldsymbol{X}\boldsymbol{\beta}$$

Thus, the standard form of a multiple regression model, comprising *n* cases and *k* variables, using the matrix notation is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{e}$$

Where:

y can be a *n* x 1 matrix of the dependent variable,

X can be a *n* x (*k* + 1) matrix containing a column of ones (for the intercept) and the *k* independent variables,

β can be a (*k* + 1) x 1 matrix containing the intercept (β_0) and the regression coefficients (β_1, β_2,…, β_k), and

e can be a *n* x 1 matrix of the error terms or residuals.

Now, recall the normal equations derived earlier by partially differentiating the sum of squared errors of a simple linear regression model:

$$Nb_0 + \sum x_i b_1 = \sum y_i$$

$$\sum x_i b_0 + \sum x_i^2 b_1 = \sum x_i y_i$$

These can be rewritten in the matrix algebra form as:

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Using the exposition above, it follows that:

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \boldsymbol{X'X} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 & . & . & . & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}$$

$$\begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \boldsymbol{X'y} = \begin{bmatrix} 1 & 1 & . & . & . & 1 \\ x_1 & x_2 & . & . & . & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

With these reformulations, the matrix version of the normal equations can be re-written as:

$$\boldsymbol{X'Xb} = \boldsymbol{X'y}$$

To solve for $\boldsymbol{b}$, we pre-multiply both sides with the inverse of $\boldsymbol{X'X}$ to get:

$$\left(\boldsymbol{X'X}\right)^{-1} \boldsymbol{X'Xb} = \left(\boldsymbol{X'X}\right)^{-1} \boldsymbol{X'y}$$

Which gives the expression for matrix of parameter estimates:

$$\boldsymbol{b} = \left(\boldsymbol{X'X}\right)^{-1} \boldsymbol{X'y}$$

This is the matrix version of the formula for least squares estimators of multiple linear equation parameters— a result of central importance in econometrics.

Let's consider a numerical example to better understand the calculations involved. We model a simplified relationship between marks scored in an econometric examination (the outcome variable) and hours spent on preparing for the test (the predictor variable). We gather the values for ten students as:

$$y_i = \{50, 20, 30, 20, 0\}$$
$$x_i = \{6, 4, 2, 0, 0\}$$

We model the relationship as:

$$y = X\beta + e \text{ with } e \sim N(0, \sigma^2 Ii)$$

We have:

$$X = \begin{bmatrix} 1 & 6 \\ 1 & 4 \\ 1 & 2 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \text{ and } y = \begin{bmatrix} 50 \\ 20 \\ 30 \\ 20 \\ 00 \end{bmatrix}$$

Then:

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6 & 4 & 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 6 \\ 1 & 4 \\ 1 & 2 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 12 \\ 12 & 56 \end{bmatrix}$$

therefore, the inverse of this matrix is:

$$(X'X)^{-1} = \begin{bmatrix} 0.41 & -0.10 \\ -0.10 & 0.04 \end{bmatrix}$$

Moreover, we also compute:

$$X'y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 6 & 4 & 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 50 \\ 20 \\ 30 \\ 20 \\ 00 \end{bmatrix}$$

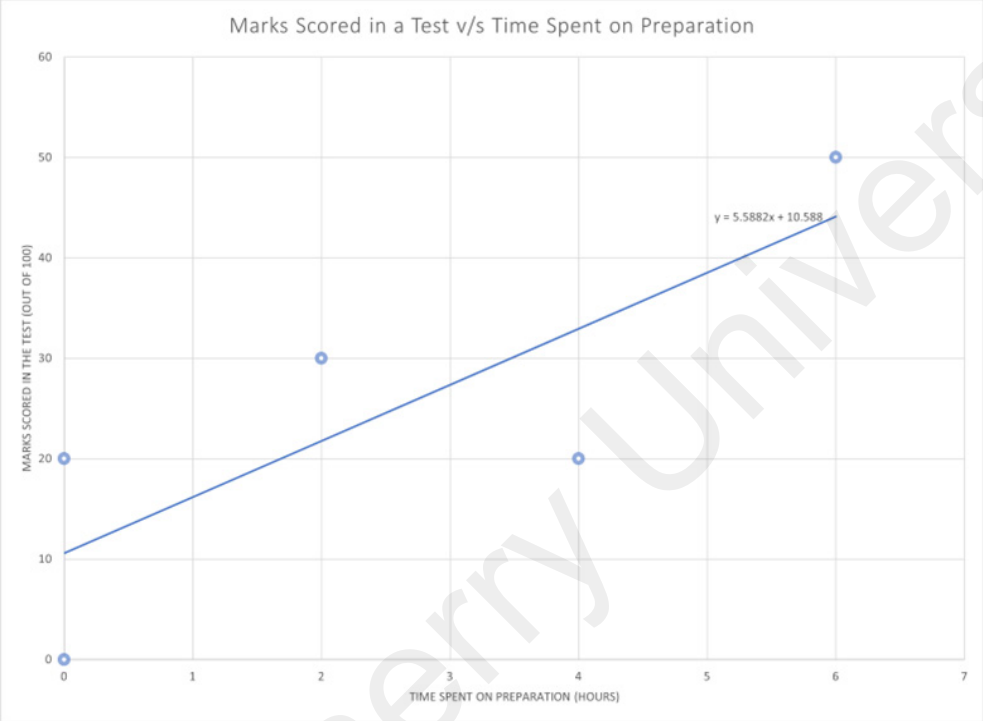
$$= \begin{bmatrix} 120 \\ 440 \end{bmatrix}$$

This gives us the final result:

$$(X'X)^{-1} X'y = \begin{bmatrix} 10.59 \\ 5.59 \end{bmatrix} = b$$

Thus, the intercept (β_0) is estimated as $b_0 = 10.59$ and the slope (β_1) is estimated as $b_1 = 5.59$. The estimated regression equation becomes: $y_i = 10.59 + 5.59x_i$, presented in the graph below.

The intercept value suggests that a student is likely to score 10.59 marks even with no time spent on preparing for the test, while every one hour spent on preparation increase the marks by 5.59, on an average.



2.2.3 OLS Estimator is BLUE

We now have the tools to show that the ordinary least squares estimator is best and unbiased among all linear estimators (**B**est **L**inear **U**nbiased **E**stimator). It refers to the properties of the OLS estimator when applied to a linear regression model under certain assumptions. A detailed explanation of each component of BLUE follows:

This is the matrix version of the formula for least squares estimators of multiple linear equation parameters— a result of central importance in econometrics.

Let's consider a numerical example to better understand the calculations involved. We model a simplified relationship between marks scored in an econometric examination (the outcome variable) and hours spent on preparing for the test (the predictor variable). We gather the values for ten students as:

B - Best: Also known as the Gauss-Markov theorem–the OLS estimator is considered the "best" because it has the smallest variance among all

linear unbiased estimators. In other words, it is the most precise or efficient estimator. Consider a simple linear model: $\boldsymbol{y} = \boldsymbol{X\beta} + \mathbf{e}$. Let's define a linear estimator other than the OLS as:

$$b^* = \sum_{i=1}^{N} a_i^* Y_i$$

Let us also assume: $a_i^* = a_i + c_i = \frac{1}{N} + c_i$. Thus,

$$b^* = \sum_{i=1}^{N} a_i^* Y_i$$

$$= \sum_{i=1}^{N} \left(\frac{1}{N} + c_i \right) Y_i$$

$$= \sum_{i=1}^{N} \frac{1}{N} Y_i + \sum_{i=1}^{N} c_i Y_i$$

$$= b + \sum_{i=1}^{N} c_i Y_i$$

Then

$$E\left[b^* \right] = E\left[b + \sum_{i=1}^{N} c_i Y_i \right]$$

$$= E\left[b \right] + \sum_{i=1}^{N} c_i E\left[Y_i \right]$$

Since, $E[b] = \beta$ (shown later), we have:

$$E\left[b^* \right] \quad \beta + \beta \sum_{i=1}^{N} c_i$$

Now let's assume that b^* is an unbiased estimator; this implies that: $\sum_{i=1}^{N} c_i = 0$. Then:

$$var\left(b^* \right) = var\left(\sum_{i=1}^{N} a_i^* Y_i \right)$$

$$= var\left(\sum_{i=1}^{N} \left(\frac{1}{N} + c_i \right) Y_i \right)$$

$$= \sum_{i=1}^{N} \left(\frac{1}{N} + c_i \right)^2 var\left(Y_i \right)$$

$$= \sigma^2 \sum_{i=1}^{N} \left(\frac{1}{N} + c_i \right)^2$$

$$= \sigma^2 \sum_{i=1}^{N} \left(\frac{1}{N^2} + \frac{2}{N} c_i + c_i^2 \right)$$

$$= \sigma^2 \left(\sum_{i=1}^{N} \frac{1}{N^2} + \frac{2}{N} \sum_{i=1}^{N} c_i + c_i^2 \Sigma_{i=1}^{N} \right)$$

$$= \sigma^2 \left(\frac{1}{N} + 0 + N c_i^2 \right)$$

$$= \frac{\sigma^2}{N} + \sigma^2 \left(N c_i^2 \right)$$

$$= var(b) + \sigma^2 \left(N c_i^2 \right)$$

This implies that variance of the newly defined estimator is greater than that of the OLS estimator:

$$\boldsymbol{var}(\boldsymbol{b}^*) > \boldsymbol{var}(\boldsymbol{b})$$

In other words, the OLS estimator has the least variance among all linear estimators.

L - Linear: The OLS estimator is a linear function of the data. This means that the estimated coefficients are a linear combination of the observed values of the dependent and independent variables. Since,

$$b = \sum_{i=1}^{N} \frac{Y_i}{N}$$

$$= \sum_{i=1}^{N} \frac{1}{N} Y_i$$

$$= \frac{1}{N} Y_1 + \frac{1}{N} Y_2 + \cdots + \frac{1}{N} Y_N$$

$$= \sum_{i=1}^{N} a_i Y_i$$

Where $\left(a_i = \frac{1}{N} \right)$ and act as 'weights' (making *b* the weighted average) thus expressing *b* as a linear combination of Y values.

U - Unbiased: An estimator is unbiased if its expected value is equal to the true value of the parameter it is estimating. In the context of OLS, this means that the expected value of the estimated coefficients is equal to the true coefficients of the population regression line.

Does the *estimated* value of the parameter equal the *actual* value of the parameter? This is equivalent of asking whether a sample statistic equals population parameter? This question cannot be answered with absolute certainty; we can only calculate the probability of this being the case. Nevertheless, a sample statistic is expected to be equal to the population parameter given that the sampling process and the estimation procedure are unbiased. Thus:

$$E[b] = E\left[\sum_{i=1}^{N} \frac{Y_i}{N}\right]$$

$$= \frac{1}{N}E[Y_1] + \frac{1}{N}E[Y_2] + \cdots + \frac{1}{N}E[Y_N]$$

$$= \frac{1}{N}\beta + \frac{1}{N}\beta + \cdots + \frac{1}{N}\beta$$

$$= \left(\frac{1}{N} \times N\right)\beta$$

$$\square E[b] = \beta$$

E - Estimator: An estimator is a rule or method used to infer the value of an unknown parameter in a statistical model based on observed data. In OLS, the estimator is used to estimate the coefficients of the linear regression model.

The BLUE properties of the OLS estimator hold under certain assumptions, including linearity, independence of errors, homoscedasticity, normality of errors, and no perfect multicollinearity. If these assumptions are met, the OLS estimator is the best (in terms of having the smallest variance) linear unbiased estimator of the coefficients of the linear regression model. We will examine in a later lesson, what happens if we relax these assumptions or if these assumptions are violated in the data collected as sample.

2.2.4 Interpretation of Coefficients in a Multiple Regression Model

The parameter estimates in an Ordinary Least Squares (OLS) regression model represent the estimated coefficients of the linear relationship between the dependent variable and the independent variables. To interpret a multiple regression model, we focus on:

- **Intercept (β_0):** The intercept is the estimated value of the dependent variable when all the independent variables are equal to zero. It

represents the baseline value of the dependent variable when the effects of all independent variables are absent.

- **Slope Coefficients ($\beta_1, \beta_2, \ldots, \beta_k$):** The slope coefficients represent the estimated change in the dependent variable for a one-unit increase in the corresponding independent variable, holding all other independent variables constant. A positive slope coefficient indicates that the dependent variable increases as the independent variable increases, while a negative slope coefficient indicates that the dependent variable decreases as the independent variable increases.
- **Significance of Coefficients:** The significance of the coefficients is typically evaluated using hypothesis testing, such as t-tests or F-tests. The p-value associated with each coefficient indicates the probability of observing the estimated coefficient value by chance, assuming that the true coefficient is zero. A small p-value (typically less than 0.05) indicates that the coefficient is statistically significant and unlikely to be zero.
- **Standard Errors:** The standard errors of the coefficients provide a measure of the precision of the estimates. A smaller standard error indicates that the estimate is more precise, while a larger standard error indicates that the estimate is less precise. The standard errors are used to calculate confidence intervals for the coefficients, which provide a range of plausible values for the true coefficient.

The parameter estimates of OLS provide information about the magnitude, direction, and significance of the relationship between the dependent variable and each independent variable, holding all other independent variables constant. The significance of the coefficients and the standard errors provide information about the reliability and precision of the estimates.

2.2.5 Summary

Multiple Linear Regression (MLR) is a fundamental statistical technique that allows us to understand and predict the behavior of one dependent variable based on the values of two or more independent variables. This method extends the concept of simple linear regression, which only considers a single predictor, to a more complex scenario where multiple factors simultaneously influence the outcome.

The core of MLR lies in its ability to model the linear relationship between the dependent variable and several independent variables. By doing so, it provides a nuanced understanding of how changes in predictor variables are associated with changes in the outcome variable. This modeling is crucial for fields such as economics, social sciences, and natural sciences, where multiple variables often interact to influence the phenomena being studied.

A key aspect of MLR is the estimation of coefficients through the Ordinary Least Squares (OLS) method. OLS is designed to minimize the sum of the squared differences between the observed values and the values predicted by the model. This approach ensures that the regression line is the best fit for the data, capturing the essential patterns without being overly influenced by outliers. The coefficients obtained from OLS estimation offer insights into the strength and direction of the relationship between each independent variable and the dependent variable, allowing for precise interpretations of how each factor affects the outcome.

Interpreting these coefficients is a critical skill in MLR. Each coefficient indicates the expected change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. This interpretation helps in understanding the individual impact of each predictor on the outcome, providing a basis for informed decision-making and policy formulation.

Moreover, MLR incorporates matrix algebra for efficient computation of regression coefficients, especially when dealing with a large number of variables. This mathematical framework simplifies the estimation process, making it more accessible and manageable.

Another important component of MLR is the assessment of the model's significance, including the statistical significance of individual coefficients, which helps in determining whether the relationships observed in the sample data are likely to exist in the broader population. The direction (positive or negative) and magnitude of the coefficients are also analyzed, offering deeper insights into the nature of the relationships between variables.

In practice, MLR is not just about fitting a model to data; it also involves diagnostic checks to ensure the model's assumptions are met, including linearity, independence, homoscedasticity, and normality of residuals.

Addressing these assumptions is crucial for the validity and reliability of the regression analysis.

2.2.6 Keywords

Multiple Linear Regression (MLR): A statistical technique for predicting the outcome of a dependent variable based on multiple independent variables.

Ordinary Least Squares (OLS): The most common method for estimating the coefficients in a regression model, minimizing the sum of squared residuals.

Coefficients: Values estimated by the regression model, indicating the magnitude and direction of the relationship between each independent variable and the dependent variable.

Dependent Variable: The outcome variable that the model aims to predict or explain, influenced by various independent variables.

Independent Variables: Predictor variables used in the model to explain variation in the dependent variable.

BLUE (Best Linear Unbiased Estimator): A property of OLS estimators that are the best (in terms of lowest variance) among all unbiased linear estimators.

Intercept: The expected value of the dependent variable when all independent variables are zero.

Slope Coefficients: Represent the expected change in the dependent variable for a one-unit increase in the corresponding independent variable, holding others constant.

Matrix Algebra: A mathematical framework used in MLR for efficient computation of regression coefficients when dealing with multiple variables.

Statistical Significance: An assessment of whether the observed relationship between variables is due to chance or represents a real effect.

2.2.7 Self-assessment Questions

1. Show that the estimated least squares line, $\hat{y_i} = b_0 + x_i b_1$, passes through the point $(\bar{x}, \bar{y})$.

2. Prove that:

$$b_0 = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i y_i}{N\Sigma x_i^2 - \left(\Sigma x_i\right)^2} = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma x_i^2 - \left(\Sigma x_i\right)^2} = \frac{\Sigma\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\Sigma\left(x_i - \bar{x}\right)^2}$$

3. Prove that:

$$\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \dfrac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i y_i}{N\Sigma x_i^2 - \left(\Sigma x_i\right)^2} \\ \dfrac{N\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N\Sigma x_i^2 - \left(\Sigma x_i\right)^2} \end{bmatrix}$$

4. Show that the sum of squared errors can be written as:

$$\sum_{i=1}^{N}\left(y_i - \beta\right)^2 = \left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}\right)'\left(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta}\right)$$

$$= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}\boldsymbol{x}'\boldsymbol{y} + \boldsymbol{\beta}^2\boldsymbol{x}'\boldsymbol{x}$$

5. Let b_{wt} be a weighted estimator with unequal weights for sample observations, Y_1, Y_2, Y_3, drawn from a normal population with a mean β and variance σ^2 such that:

$$b_{wt} = \frac{1}{2}Y_1 + \frac{1}{3}Y_2 + \frac{1}{6}Y_3$$

Show that b_{wt} is a linear, unbiased estimator.

6. For a simple linear statistical estimation model, show that sum of errors will always sum to zero. That is:

$$\sum_{i=1}^{N}\left(y_i - b\right) = 0$$

7. Randomly selected ten households in a city had the following income and expenditure on clothes (in rupees thousands):

Income	30	20	40	33	13	15	38	26	43	35
Expenditure	9	7	11	8	4	5	10	8	10	9

Estimate the expenditure–income regression line. By what amount does expenditure on clothes increase if the income increases by Rs.1000?

8. Estimate the demand function of cotton candy from the price and quantity bought data below:

Price	16	18	14	10	12	7	10	6	9	10
Quantity	150	100	200	275	250	300	280	325	260	280

Forecast the demand when the price is Rs.15.

9. Assuming a linear relation Y between X (outcome variable) and (predictor variable), for a sample of 11 observations:

$$\overline{X} = 520.20$$

$$\overline{Y} = 220.80$$

$$\sum X_i^2 = 3134543$$

$$\sum X_i Y_i = 1296836$$

$$\sum Y_I^2 = 539512$$

Estimate the linear regression line and interpret the parameter estimates.

10. Following is the data for Central Goods and Services Tax (CGST) collection (in rupees thousand crores) for the last ten years:

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
CGST	150	160	100	230	250	300	80	140	235	320

Using OLS method, estimate the slope of the CGST regression line.

2.2.8 References

1. **Introduction to Linear Regression Analysis** by Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. This textbook is a staple in regression analysis, offering a thorough overview of the subject. It covers the basics of simple linear regression and extends to multiple regression, diagnostics, and the application of regression analysis in various fields. Its clear explanations and practical approach make it an excellent choice for students.
2. **Applied Linear Statistical Models** by Michael Kutner, Christopher Nachtsheim, John Neter, and William Li. This comprehensive text covers a wide range of topics, including simple and multiple

linear regression, analysis of variance (ANOVA), and non-linear models. It's known for its applied approach, with numerous real-world examples and case studies that help students understand the practical applications of regression analysis.

3. **Applied Regression Analysis and Generalized Linear Models** by John Fox. John Fox's textbook is geared towards students and researchers looking to understand the theory and application of regression analysis, including linear, multiple linear, and generalized linear models. The book emphasizes the social science perspective, making it particularly useful for students in fields like economics, sociology, and political science.
4. **Econometric Analysis** by William H. Greene. Greene's textbook is a classic in the field of econometrics, providing comprehensive coverage of theoretical and applied aspects of econometric methods, including multiple regression analysis. It's well-suited for students with an interest in the economic applications of regression and those looking for a deeper understanding of the statistical foundations.
5. **Statistics and Data Analysis for Financial Engineering** by David Ruppert. This book is designed for students interested in the financial applications of statistical methods, including regression analysis. It covers a broad range of topics, from basic statistical concepts to more complex models like multiple regression and time series analysis, all with a focus on applications in financial engineering.

UNIT – III : THE PROBLEM OF INFERENCE

Lesson 3.1 – Evaluation of the Least Squares Estimates

Structure

3.1.1 Evaluating the Reliability of the Estimates

The reliability of least squares estimates refers to the consistency and precision of the estimated regression coefficients in a linear regression model. It reflects the degree of confidence that these estimates are accurate and measure the true underlying relationship between the dependent and independent variables in a linear regression model. When we speak of the reliability of least squares estimates, we are essentially asking:

- **How much do our estimates change if we have slightly different data?** If small changes in our data lead to drastically different estimates, we wouldn't consider those estimates reliable.
- **How well do our estimates capture the true underlying parameters of the model?** Even with consistent estimates, they might not accurately reflect the real-world relationships we're trying to model.

This concept is crucial in statistics and econometrics because it influences the interpretability and applicability of regression analysis results. To assess and evaluate the reliability of least squares estimates, several criteria and methods are employed, including the coefficient of determination (R^2), standard errors of the estimates, confidence intervals, hypothesis testing, and the analysis of residuals.

3.1.2 Criteria for Assessing the Reliability

Reliability of a regression model can be judged on three set of criteria: *a priori* theoretical justification of the model, *first order* statistical tests of the estimates, and *second order* econometric considerations for the sample data on which the regression is run.

Even before a regression model is constructed, theory must be consulted for establishing the nature and direction (possibly, relative strength) of relationship between the variables. It begins with proper identification of the outcome and predictor variables and then postulating a relationship between them. This step is followed by constructing a mathematical model of the theoretical relationship. At this stage, theory must guide as to the exact form of the relationship, relative strength and the direction of influence among the variables. Only a theoretically sound model can have any meaning attached to its estimates.

The statistical model establishes the data collection strategy an the econometric model then estimates the model parameters. The primary criteria for assessing the reliability of econometric method used (the estimator), are unbiasedness, consistency, and efficiency.

Unbiasedness: An estimator is unbiased if its expected value is equal to the true population parameter. In the case of least squares estimates, the Gauss-Markov theorem states that, under certain assumptions (linearity, independence, homoscedasticity, and exogeneity), the least squares estimators are unbiased and have the smallest variance among all linear unbiased estimators. This means that, on average, the least squares estimates will be equal to the true population values.

Consistency: An estimator is consistent if it converges in probability to the true population parameter as the sample size increases. In other words, as more data becomes available, the estimator becomes more accurate and reliable. Least squares estimators are consistent under the assumptions of the classical linear regression model.

Efficiency: An estimator is efficient if it has the smallest variance among all unbiased estimators. The Gauss-Markov theorem states that least squares estimators are efficient within the class of linear unbiased estimators. However, they may not be the most efficient estimators overall, especially if some of the model assumptions are violated.

We already know that the ordinary least squares estimator is BLUE under the assumptions of the model. These assumptions must be checked for in individual sample regression model estimates, both graphically (using diagnostic plots such as residual and normal probability plots) and numerically (for detecting the presence of multicollinearity, heteroscedasticity, and autocorrelation). We will discuss these assumptions and consequences of their violations as well their detection in the data and resolution of the problem in later lessons. For now, assuming the *second order* econometric conditions hold, we will understand the statistical procedures to assess the reliability of the estimates using the *first order* tests.

Reliable estimates are those that are close to the true population values and have small standard errors, indicating that they would not vary substantially if the sampling process were repeated. The least squares method aims to minimize the sum of the squared differences between the observed values and the values predicted by the linear model. Given a dataset with a dependent variable and an independent variable X, the linear regression model can be expressed as:

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

where β_0 is the intercept, β_1 is the slope of the regression line, and ϵ represents the error term. The least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are calculated to minimize the sum of squared residuals (*SSR*):

$$SSR = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i \right)^2$$

where,

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

are the predicted values. Before assessing the individual parameter estimates, we assess how well does the proposed linear regression model fit the data. This is done by comparing the predicted values of the regression model with an alternate model.

Suppose a sales manager wants to predict the sales for next month. How does he do that? In absence of any information–factors that affects sales, the exact form of mathematical relationship those factors have with sales–the best he can do is to find out the average sales of the past few months and take that number as the *expected* sales for the coming month. Finding out the average, thus, can be seen as a *model* for making prediction rather

than just a number. This is the most basic of all–the simplest of all possible models to abstract reality. Any other model, then, needs to perform better than the average to be considered as a good model or even a model itself.

3.1.3 Evaluating the Overall Goodness of Fit

A linear regression model is an improvement over the finding out the average model if we can show that the values of the outcome variable predicted by it for different values of the predictor variables is closer to the actual observed values of the outcome variable. We do not expect, in general, the predicted values to be exact matches to the observed values–be it either the average ($\bar{Y}$) or the linear regression estimate ($\hat{Y}$). The model whose predicted values are closer to the observed ones will be the better model. Thus, we need to essentially compare ($Y - \bar{Y}$) with ($Y - \hat{Y}$). This can be visualized in the figure below:

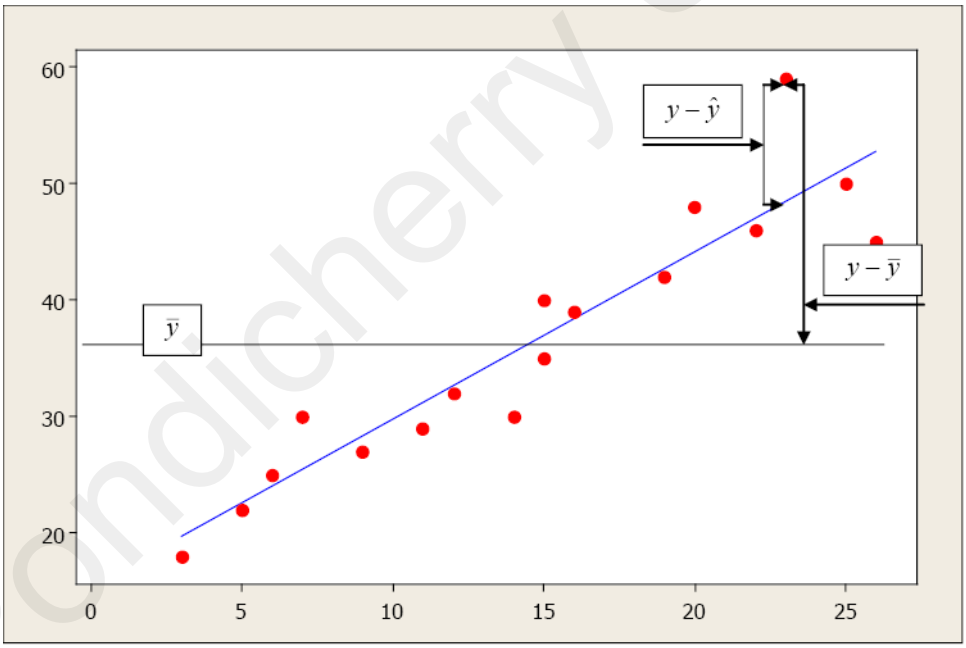


Figure 1: Assessing the Fit of the Regression Line
(This Photo by Unknown Author is licensed under CC BY-SA-NC)

In the figure above, the (red-colored) dots are the observed values of y, the horizontal line is the average ($\bar{y}$) of these observed values, and the (blue-colored) upward sloping straight line is the regression line on which lie the predicted values ($\hat{y}$) of the outcome variable. So, for all values of the predictor variable (x), the *average model* makes the same prediction ($\bar{y}$) and the difference between the two is ($y - \bar{y}$). But the regression model makes different predictions for different values of ; sometimes matching the exact observed value of y (dots overlapping the regression line), sometimes

overestimating (dots below the regression line) or underestimating (dots above the regression line) them.

Thus, for any particular value of x, the difference between the corresponding observed value of $(y - \bar{y})$; and the predicted value given by the *naïve* model is (); it can be thought of as *total deviation* from the expected value ($E[y] = \bar{y}$). Now, we can ask ourselves this question: what explains the deviation of the observed value from the expected value? The corresponding regression estimate is able to account for a part (if not all) of this deviation since it lies between the mean and the observed value. Thus, $(\hat{y} - \bar{y})$ can be thought of as the *explained deviation*. And the rest, the gap between the observed value and the regression estimate $(y - \bar{y})$) is the *unexplained variation* (or residual, or simply, the error) of the observed value from the expected value of the outcome variable. Thus,

$$\left[Total\ Deviation\right] = \left[Explained\ Deviation\right] + \left[Unexplained\ Deviation\right]$$

$$\left(y - \bar{y}\right) = \left(\hat{y} - \bar{y}\right) + \left(y - \hat{y}\right)$$

These deviations are summed for each value of x and is called variously as:

$$TSS = \sum_{1}^{N}\left(Y_i - \bar{Y}\right)^2 \left[Total\ Sum\ of\ Squared\ Errors\right]$$

$$ESS = \sum_{1}^{N}\left(\hat{Y} - \bar{Y}\right)^2 \left[Explained\ Sum\ of\ Squared\ Errors\right]$$

$$RSS = \sum_{1}^{N}\left(Y - \hat{Y}\right)^2 \left[Residual\ Sum\ of\ Squared\ Errors\right]$$

Thus, for a regression model:

$$TSS = ESS + RSS$$

$$\sum_{1}^{N}\left(Y_i - \bar{Y}\right)^2 = \sum_{1}^{N}\left(\hat{Y} - \bar{Y}\right)^2 + \sum_{1}^{N}\left(Y - \hat{Y}\right)^2$$

Looked at this way, the higher the proportion of variation explained by a model, the better it is compared to the *naïve* model. The proportion of variation explained by the model is:

$$\frac{ESS}{TSS} = \frac{\sum_{1}^{N}\left(\hat{Y} - \bar{Y}\right)^2}{\sum_{1}^{N}\left(Y_i - \bar{Y}\right)^2}$$

We define this ratio as the *coefficient of determination*:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Now, in a linear regression model, there are three possibilities:

1. The variations in the observed values is completely explained by the regression estimates. In this case $RSS = 0$, therefore $R^2 = 1$.
2. The variations in the observed values are not at all explained by the regression estimate. In this case $RSS = 1$, therefore $R^2 = 0$.
3. Variations in the observed values are partially explained by the regression estimate. In this case $RSS < TSS$, that is, $\frac{RSS}{TSS} < 1$, and therefore the value of R^2 will be less than 1 but greater than zero.

In summary, the values of R^2 lie between zero and one:

$$0 \leq R^2 \leq 1$$

The closer the value of R^2 to 1, the better the regression model estimates.

Consider the following dataset of GDP and direct tax collection (in billions of rupees). We model the relationship as linear, estimate the regression line and assess how well does it explain the variation in the sample data:

Y_i **(Direct Tax)**	X_i **(Real GDP)**	$\hat{Y}_i$	***ESS***	***RSS***	***TSS***
3.5	16	3.45	0.20	0.00	0.21
3.2	14	3.15	0.02	0.00	0.03
3.0	12	2.85	0.02	0.02	0.04
2.6	11	2.70	0.09	0.01	0.10
2.9	12	2.85	0.02	0.00	0.03
3.3	15	3.30	0.09	0.00	0.09
2.7	13	3.00	0.00	0.09	0.09
2.8	11	2.70	0.09	0.01	0.10
3.0		**TOTAL**	**0.54**	**0.14**	**0.68**

The coefficient of determination is:

$$R^2 = \frac{ESS}{TSS} = \frac{0.54}{0.68} = 0.79$$

This implies that the regression model accounts for 79% of variations in the observed values. In other words, variations in the predictor variable, *X*, is able to explain 79% of the variations in the outcome variable, *Y*. This leaves 21% variations in direct tax collections unexplained by changes in the GDP of the economy. This is a 'good' fit; although good being a relative word, the model fit is more towards the ideal 100% than far from it.

3.1.4 Testing the Significance of the Parameter Estimates

The regression line is estimated on (representative) sample drawn from the population. Since the population regression line is unknown, the sample regression line can only be an estimate. Therefore, the parameter estimates could well be chance values, i.e., we need to test whether they are truly non-zero or not. As statistical tests go, we specify the alternate and the null hypothesis for the slope of the sample regression line as:

$$H_0: \beta_1 = \hat{\beta}_1$$

$$H_1: \beta_1 \neq \hat{\beta}_1$$

But to test these hypotheses, we need to know about the characteristics of the estimator as well as its probability distribution.

We begin with descriptive statistics. Since β is estimated from a sampling procedure where the *x* values remain the same from one sample to another, theoretically, β is a random variable. The expected value of a random variable is its mean. We will derive an expression for the expected value of the least squares estimator. Recall from earlier, the slope of the sample regression line is:

$$\hat{\beta}_1 = \frac{\Sigma\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\Sigma\left(X_i - \bar{X}\right)^2}$$

This can be shortened to:

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^{\,2}}$$

Where:

$$y_i = \left(Y_i - \bar{Y}\right) x_i = \left(X_i - \bar{X}\right)$$

Now, the estimated regression slope can be rewritten as:

$$\hat{\beta}_1 = \sum w_i y_i$$

Where:

$$w_i = \frac{x_i}{\Sigma x_i^2}$$

Thus, estimated slope of the regression line can be expressed as the weighted sum of squared deviations of the outcome variable from its mean. These weights, w_i, remain fixed in repeated sampling and possess the following properties:

$$\sum w_i = 0$$

$$\sum w_i^2 = \frac{1}{\Sigma x_i^2}$$

$$\sum w_i x_i = \sum w_i X_i = 1$$

The above exposition implies that:

$$\hat{\beta}_1 = \sum w_i Y_i$$

Then substituting the value of Y_i from a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

We get:

$$\hat{\beta}_1 = \beta_0 \sum w_i + \beta_1 \sum w_i X_i + w_i e_i$$

$$= \beta_1 + w_i e_i$$

Taking the expected value of the slope estimator, we get:

$$E\left[\hat{\beta}_1\right] = \beta_1 + w_i E\left[e_i\right]$$

Since, the error term is expected on an average to be zero, i.e., $E\left[e_i\right] = 0$, this gives:

$$E\left[\hat{\beta}_1\right] = \beta_1$$

Thus, the mean of the ordinary least squares estimate of the sample regression line is equal to the true slope of the population regression line. The variance of the slope estimator can be found out as:

$$var\left(\hat{\beta}_1\right) = E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right]$$

$$= E\left[\left(\sum w_i e_i\right)^2\right]$$

$$= E[\left(w_1 e_1 + w_2 e_2 + \ldots + w_N e_N\right)^2]$$

$$= E\left[w_1^2 e_1^2 + w_2^2 e_2^2 + \ldots + w_N^2 e_N^2 + 2w_1 w_j e_i e_j + \ldots\right]$$

The remaining terms of which are cross-products of $w_i w_j e_i e_j$ for all $(i < j)$. And, since:

$$var\left(e_i\right) = E\left[e_i^2\right] = \sigma^2 \; for\, all\, i$$

$$cov\left(e_i, e_j\right) = E\left[e_i e_j\right] = 0 \; for\, i \neq j$$

This simply reduces to:

$$var\left(\hat{\beta}_1\right) = E\left[\left(\sum w_i e_i\right)^2\right]$$

$$= \sigma^2 \sum w_i^2$$

$$= \frac{\sigma^2}{\Sigma x_i^2}$$

The mean of the intercept estimate, $\hat{\beta}_0$, can be found as:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

$$= \beta_0 + \beta_1 \overline{X} + \overline{e} - \hat{\beta}_0 \overline{X}$$

$$= \beta_0 - \left(\hat{\beta}_1 - \beta_1\right)\overline{X} + \overline{e}$$

Since, $E\left[\hat{\beta}_1\right] = \beta_1$ and $E\left[e_i\right] = 0$, taking expectations of both sides gives:

$$E\left[\hat{\beta}_0\right] = \beta_0$$

Then the variance of the intercept estimate can be found out as:

$$var\left(\hat{\beta}_0\right) = E\left[\left(\hat{\beta}_0 - \beta_0\right)^2\right]$$

$$= \overline{X}^2 E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] + E\left[\overline{e}^2\right] - 2\overline{X}E\left[\left(\hat{\beta}_1 - \beta_1\right)\overline{e}\right]$$

Now,

$$E\left[\left(\hat{\beta}_1-\beta_1\right)^2\right]=\frac{\sigma^2}{\Sigma x^2}$$

And

$$E\left[\bar{e}^2\right]=\frac{\sigma^2}{N}$$

Thus:

$$E\left[\left(\hat{\beta}_1-\beta_1\right)\bar{e}\right]=E\left[\left(\Sigma w_i e_i\right)\ \left(\frac{1}{N}\Sigma e_i\right)\right]$$

$$=E\left[\frac{1}{N}\left(\Sigma w_i e_i^2+cross-product\ terms\ in\ e_i e_j\right)\right]$$

$$=\frac{1}{N}\sigma^2\Sigma w_i$$

$$=0$$

Therefore:

$$var\left(\hat{\beta}_0\right)=\sigma^2\left[\frac{1}{N}+\frac{\bar{X}^2}{\Sigma x^2}\right]$$

And, finally, the error term–itself a random variable– can be expressed as:

$$e_0=Y_0-\hat{Y}_0=-\left(\hat{\beta}_1-\beta_1\right)x_0+\dot{o}_0-\bar{o}$$

If we square both sides of this expression and take expectations, the expectations of the cross-product terms vanish. Also, since ȯ's are independent and from an earlier result, $E\left[\left(\hat{\beta}_1-\beta_1\right)\bar{e}\right]=0$, we get:

$$var\left(e_0\right)=E\left(\dot{o}_0^2\right)+E\left[\bar{o}^2\right]+x_0^2E\left[\left(\hat{\beta}_1-\beta_1\right)^2\right]$$

$$=\sigma^2\left[1+\frac{1}{N}+\frac{x_0^2}{\Sigma x^2}\right]$$

Again, since $\boldsymbol{E}[\boldsymbol{e}]=0$, the variance–covariance matrix of the error term can be represented as the expected value of the matrix:

$$\boldsymbol{ee'}=\begin{bmatrix}e_1\\e_2\\\vdots\\e_N\end{bmatrix}\begin{bmatrix}e_1 & e_2 & \cdots & e_N\end{bmatrix}$$

$$= \begin{bmatrix} e_1^2 & e_1e_2 & \cdots & e_1e_N \\ e_1e_2 & e_2^2 & \cdots & e_2e_N \\ \vdots & \vdots & \vdots & \vdots \\ e_Ne_1 & e_Ne_2 & \cdots & e_N^2 \end{bmatrix}$$

Using the rule for mathematical expectation of a matrix, we get:

$$\boldsymbol{E}\left[\boldsymbol{ee'}\right] = \begin{bmatrix} E(e_1^2) & E\left(e_1e_2\right) & \cdots & E\left(e_1e_N\right) \\ E\left(e_1e_2\right) & E\left(e_2^2\right) & \cdots & E\left(e_2e_N\right) \\ \vdots & \vdots & \vdots & \vdots \\ E\left(e_Ne_1\right) & E\left(e_Ne_2\right) & \cdots & E\left(e_N^2\right) \end{bmatrix}$$

$$= \begin{bmatrix} var\left(e_1\right) & covar\left(e_1,e_2\right) & \cdots & covar\left(e_1,e_N\right) \\ covar\left(e_1,e_2\right) & var\left(e_2\right) & \cdots & covar\left(e_2,e_N\right) \\ \vdots & \vdots & \vdots & \vdots \\ covar(e_N,e_1) & covar\left(e_N,e_2\right) & \cdots & var\left(e_N\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

Therefore, the variance–covariance matrix of the error term (in short, the covariance matrix of the random vector $\boldsymbol{e}$) is:

$$cov\left(\boldsymbol{e}\right) = E\left[\boldsymbol{ee'}\right] = \sigma^2 \boldsymbol{I}_N$$

Now, since by assumption, the random error term $\boldsymbol{e} \sim N\left(0, \sigma^2 \boldsymbol{I}_N\right)$, it can be easily shown that the parameter estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, are also distributed normally. We are now set to perform the test of hypothesis proposed earlier. But, before we begin, let's summarize the results that we have deduced so far:

$$Y \sim N\left(X\beta, \sigma^2 I_N\right)$$

$$e \sim N\left(0, \sigma^2 I_N\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{N} + \frac{\bar{X}^2}{\Sigma x^2}\right]\right)$$

$$\hat{\beta}_1 \quad N\left(\beta_1 \frac{\sigma^2}{\Sigma x_i^2}\right)$$

Assuming the population variance, σ^2, is known or else the sample size is sufficiently large ($N \geq 30$), the ordinary least squares estimates can be standardized following the Z-distribution formula as:

$$Z_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim N(0,1)$$

$$Z_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$$

Using these expressions, we can calculate the probability of the repeated sampling producing the value $\hat{\beta}_0$ from a population that has the hypothesized value (k) as true β_0. Working out in this way, the null and alternate hypotheses are written as:

$$H_0 : \beta_0 = k$$

$$H_1 : \beta_0 \neq k$$

The hypothesized value k could be suggested by economic theory or may have been obtained in previous studies. The probability of observing $\hat{\beta}_0$, that is $P\left(Z_{\hat{\beta}_0}\right)$, is then compared to a pre-specified level of significance, usually taken as 5% ($\alpha = 0.05$), with the decision criteria being rejection of null hypothesis if $P\left(Z_{\hat{\beta}_0}\right) < \alpha$. This implies that the probability of getting $\hat{\beta}_0$ by chance is extremely low and therefore, highly unlikely that the true value (β_0) of the population parameter is different from the estimated value ($\hat{\beta}_0$) of the sample statistic.

But if the theory does not guide and there are no previous studies to look into, the hypothesis then reduces to simply checking whether the data gives evidence that the parameter has non-zero value, i.e., the associated variable is *significant*. This is the customary two-tailed test usually conducted in econometric studies:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

In cases where the objective is to check whether the true parameter is non-zero, calculating *p*-value for the sample estimate reduces to dividing it with the standard deviation:

$$Z_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 0}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}$$

An alternative to the *p*-value approach towards hypothesis testing is the construction of confidence intervals of values around the hypothesized true value of the parameter and check whether the estimated value falls within the interval so constructed—if it does, the null hypothesis is accepted, if it does not, the alternate hypothesis is accepted.

The width of the confidence interval depends on the level of significance chosen. For the commonly used $\alpha = 0.05$, the confidence interval can be constructed using the *Z*-value that corresponds to $1-\alpha(=0.95)$ obtained from the standard normal table as 1.96; using this, we can construct a confidence interval around the hypothesized value of the slope of the population regression line such that:

$$P\left[\left(\hat{\beta}_1 - 1.96\sigma_{\hat{\beta}_1}\right) \le \beta_1 \le \left(\hat{\beta}_1 + 1.96\sigma_{\hat{\beta}_1}\right)\right] = 0.95$$

This 95% level of confidence implies that in repeated sampling, about 95% of samples will have their $\hat{\beta}_1$ value lying between the confidence interval constructed above. In general, the confidence interval around the hypothesized true value at any level of confidence can be constructed as follows:

$$\hat{\beta}_1 \pm Z_{critical} * \sigma_{\hat{\beta}_1}$$

Where $Z_{critical}$ is the *Z*-value for the chosen level of confidence $(1-\alpha)$. But one must keep in mind that for the estimated interval constructed for a particular sample, may or may contain the true population parameter. Let's take an example: suppose $\hat{\beta}_1 = 33.21$, $\sigma_{\hat{\beta}_1} = 36$ and we want to test the hypothesis:

$$H_0 : \beta_1 = 30$$

$$H_1 : \beta_1 \neq 30$$

We get the corresponding *Z*-value as:

$$Z_{\hat{\beta}_1} = \frac{33.21 - 30}{36} = 0.089$$

Since, $0.089 < 1.96$, that is $Z_{\hat{\beta}_1} < Z_{critical}$, we reject the null hypothesis and accept the true population slope to be 30 with 95% level of confidence.

Consider another example: from a sample of 850 respondents, expenditure on food as a function of monthly income is estimated as:

$$\hat{Y} = 75 + 4.00X$$

Where, $\hat{Y}$ is the estimated expenditure on food and X is the monthly income. If the standard errors of the intercept and the slope estimates are:

$$\hat{\sigma}_{\hat{\beta}_0} = 25 \text{ and } \hat{\sigma}_{\hat{\beta}_1} = 1.25$$

And we want to test the hypothesis:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Then we compute the Z-value of the slope estimate as:

$$Z_{\hat{\beta}_1} = \frac{4.00}{1.25} = 3.33$$

Since, $3.33 > 1.96$, that is $Z_{\hat{\beta}_1} > Z_{critical}$, we reject the null hypothesis and accept the true population slope to be 4.00 with 95% level of confidence.

The above procedure is possible only if the population variance is known. But, in practice, it is not so because:

$$\sigma_{\hat{\beta}_0} = f\left(\sigma^2\right)$$

$$\sigma_{\hat{\beta}_1} = f\left(\sigma^2\right)$$

And the true population variance (σ_{δ}^2) is not known because since the error terms are unobservable. In such cases, while Z transformation cannot be performed, it is possible to replace σ_{δ}^2 with its unbiased estimator:

$$\hat{\sigma}_{\delta}^2 = \frac{\Sigma e^2}{N-1}$$

$$= \frac{\Sigma_1^N \left(Y_i - \hat{\beta}\right)^2}{N-1}$$

$$= \frac{\Sigma_1^N \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right)}{N-1}$$

Replacing the denominator with this unbiased estimator renders the transformation process follow a t-distribution with $N-k$ degrees of freedom, where k is the number of parameters in the model, including the intercept. Thus, transforming the slope estimate, we have:

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}}}$$

Again, the critical values of t_c can be found from the table of t-distribution for corresponding degrees of freedom. The confidence interval can be constructed in the usual manner with symmetrical values of t_c forming the lower and upper boundaries:

$$P\left[-t_c \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}}} \leq t_c\right] = 1-\alpha$$

$$= P\left[(\hat{\beta}_1 - t_c * \hat{\sigma}_{\hat{\beta}}) \leq \beta_1 \leq \left(\hat{\beta}_1 + t_c * \hat{\sigma}_{\hat{\beta}}\right)\right]$$

For the commonly used level of significance ($\alpha = 0.05$), the 95% confidence interval can be built using $t_c = 2.228$ for $N-1$ degrees of freedom. Consider an example: from a survey of 20 respondents, the consumption function is estimated as:

$$\hat{C} = 120 + 0.80Y$$

The standard errors of the two estimates are:

$$\hat{\sigma}_{\hat{\beta}_0} = 75.7 \text{ and } \hat{\sigma}_{\hat{\beta}_1} = 0.25$$

To check whether the slope is non-zero, we formulate the hypothesis:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_1 : \hat{\beta}_1 \neq 0$$

We transform the slope estimate as:

$$t_{\hat{\beta}_1} = \frac{0.80}{0.25} = 3.2$$

And compare this with the critical value $t_{critical}$. From the table of values of the t-distribution, we find the critical values for $(N-k) = (20-2) = 18$ degrees of freedom as:

$$t_{0.025} = \pm 2.10$$

Clearly, $t_{\hat{\beta}_1} > +2.10$, that is, the estimated value of the slope lies beyond the upper value of $t_{critical}$, therefore, we reject the null hypothesis and assert that the estimated value of the slope is statistically significant.

Consider another example. For a brand of luxury bathing soaps, its sales as a function of advertisement expenditure is estimated as:

$$\hat{Y} = 40 + 4.25X$$

Where, $\hat{Y}$ is the estimated sales and X is the advertisement expenditure. The standard errors of the intercept and the slope estimates are found to be:

$$\hat{\sigma}_{\hat{\beta}_0} = 10.2 \; and \; \hat{\sigma}_{\hat{\beta}_1} = 0.75$$

If we want to test the hypothesis:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_1 : \hat{\beta}_1 \neq 0$$

Then, we transform the slope estimate to get:

$$t_{\hat{\beta}_1} = \frac{4.25}{0.75} = 5.667$$

Now, for a 11-month sample data, the degrees of freedom are (11 – 2 = 9) and the critical values of the *t*-distribution are:

$$t_{0.025} = \pm 2.201$$

Thus,

$$\left(t_{\hat{\beta}_1} = 5.667\right) \geq \left(t_{0.025} = 2.201\right)$$

Therefore, we take the slope of the sales function as statistically significant.

3.1.5 Testing the Equality of Two Regression Coefficients

Suppose an online grocery shopping company has two service verticals: Vertical A specializes in home deliveries within 30 minutes of placing an order online, while Vertical B delivers purchased items the next day. The company wants to know whether the two verticals have the same average weekend sales of, say, processed cheese. This is an example of a problem where we want to know whether two populations have the same mean parameter.

In order to find out this, the sales manager collects sales data for the last 20 weekends. Usually, such data are not distributed normally, but techniques exist to transform non-normal data in such a way that the transformed dataset follows normal distribution. One such technique is to take logarithms of the data and use it instead of the original values.

For this example, let's assume that the logarithm of the weekend sales data is indeed normally distributed. Also, we assume that while the average sales may vary between the two verticals, they have the same standard deviation. Assuming common variance eases the exposition a little without losing much of the general case. Thus, we suppose:

$$Y_A = log\left(\text{vertical A cheese sales}\right) \sim N\left(\beta_A, \sigma^2\right)$$

$$Y_B = log\left(\text{vertical B cheese sales}\right) \sim N\left(\beta_B, \sigma^2\right)$$

Thus, the manager wants to test the following hypotheses:

$$H_0 : \beta_A = \beta_B$$

$$H_1 : \beta_A \neq \beta_B$$

We also make the following assumptions: (i) the (log of) weekend sales each vertical is independent of each other, and (ii) the (log of) weekend sales of each vertical is independent from weekend to weekend. The second point needs more clarity: we assume that sale of cheese during one weekend is independent of the sale of cheese during any other weekend—past or future. And that this is true for both the verticals. The complete model specification is:

$$Y_{Ai} = \beta_A + e_{Ai} \quad Y_{Bi} = \beta_B + e_{Bi}$$

$$e_{Ai} \sim N\left(0, \sigma^2\right) \quad e_{Bi} \sim N\left(0, \sigma^2\right)$$

$$cov\left(e_{Ai}, e_{Aj}\right) = 0 \quad cov\left(e_{Bi}, e_{Bj}\right) = 0 \quad i \neq j$$

$$cov\left(e_{Ai}, e_{Bj}\right) = 0$$

Then the ordinary least squares estimators for the two slopes are:

$$b_A = \frac{\Sigma_{i=1}^{N} Y_{Ai}}{N} \quad N\left(\beta_A, \frac{\sigma^2}{N}\right)$$

$$b_B = \frac{\Sigma_{i=1}^{N} Y_{Bi}}{N} \quad N\left(\beta_B, \frac{\sigma^2}{N}\right)$$

The 'pooled' estimator for the common variance of the two populations is:

$$\hat{\sigma}^2 = \frac{\Sigma_{i=1}^{N}\left[\left(Y_{Ai} - b_A\right)^2 + \left(Y_{Bi} - b_B\right)^2\right]}{2\left(N-1\right)}$$

It can be shown that the random variable defined as:

$$d = b_A - b_B$$

follows the *t*-distribution. Hence, the appropriate test statistic for testing the null hypothesis is given by:

$$t_d = \frac{d}{\left(\frac{2\hat{\sigma}^2}{N}\right)^{\frac{1}{2}}} \sim t_{2(N-1)}$$

The null hypothesis being rejected if $|t| \geq t_{critical}$. The critical value is obtained from the table of *t*-distribution values for $2(N-1)$ degrees of freedom.

Let us see this through by considering the following example of weekend sales figures for cheese sold through two separate verticals of the online grocery shopping company. The logarithm values of weekend sales for twenty weeks of both, Vertical A and Vertical B, are given in the following table:

Weekend	Vertical A	Vertical B
1	6.15698	6.62274
2	6.10702	7.09506
3	6.28413	7.19068
4	6.29711	6.81235
5	6.08677	6.80128
6	6.16331	6. 51767
7	6.57368	7.21671
8	6.46459	7.00307
9	6.39693	7.23634
10	6.15910	6.99485
11	6.38856	7.13090
12	5.92158	6.60123
13	5.97126	6.25958
14	5.94017	6.24804

Weekend	Vertical A	Vertical B
15	5.93754	6.47080
16	6.08904	6.95940
17	6.32615	6.87626
18	6.13556	6.55678
19	6.15910	6.31173
20	6.04263	6.78672

From the data in the table the mean sales and pooled variance are estimated:

$$b_A = \sum_{i=1}^{N} \frac{y_{Ai}}{N} = 6.18006$$

$$b_B = \sum_{i=1}^{N} \frac{y_{Bi}}{N} = 6.78461$$

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \frac{\left(y_{Ai} - b_A\right)^2 + \left(y_{Bi} - b_B\right)^2}{2\left(N-1\right)} = 0.06885$$

The value of the *t*-statistic comes out to be:

$$t_d = \frac{d}{\left(\frac{2\hat{\sigma}^2}{N}\right)^{\frac{1}{2}}} = \frac{6.18006 - 6.78461}{\sqrt{\frac{2\left(0.06885\right)}{20}}} = -7.28$$

The degrees of freedom is calculated to:

$$2\left(N-1\right) = 2\left(20-1\right) = 38$$

At α = 0.05, the critical value turns out to be: $t_{38} = 2.024$ (calculated using statistical software SPSS). Since,

$$\left|t\right| > t_{critical} \text{ or } \left|-7.28\right| > 2.024$$

we reject the null hypothesis of equality of mean sales; the two verticals indeed have different mean weekend sales of cheese.

3.1.6 Testing the Overall Significance of the Sample Regression

The *F*-statistic is the measure of overall fit of the linear regression, or for that matter, any statistical model. The *F*-test and *R*-squared (R^2) both

provide information about the fit of the regression model, but they do so in different ways and serve different purposes.

R-squared is a measure of the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit. However, R-squared has some limitations. For example, it always increases as one adds more predictors to the model, even if those predictors are not statistically significant or do not improve the model's predictive power.

The *F*-test, on the other hand, tests the null hypothesis that all the regression coefficients are simultaneously equal to zero. It provides a *p*-value that indicates whether the overall regression model is statistically significant. The *F*-test considers both the explained variance (like *R*-squared) and the unexplained variance. Unlike *R*-squared, the *F*-test can help determine whether adding more predictors to the model significantly improves the fit, or whether the apparent improvement could be due to chance.

While *R*-squared is a descriptive measure of the fit of the model, the *F*-test is an inferential test. Both can be useful in different contexts, and they complement each other in understanding the regression model's performance.

Thus, so far we have developed procedures to determine whether any individual predictor variable has statistically significant impact on the outcome variable by testing the estimate of the associated parameter for having a non-zero value–one at a time–we now extend this procedure to test whether *all* parameters have non-zero values *simultaneously*. This is a joint test of significance for parameters of all explanatory variables included in the model.

A general linear econometric model having K unknown coefficients and $(K - 1)$ explanatory variables can be written as:

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \ldots + x_{iK}\beta_K + e_i$$

Which can be succinctly written in the matrix algebra form as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

To check the overall fit of the model, in other words, to check whether all predictor variables included in the model do indeed explain the variance in the outcome model, we hypothesize:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_K = 0$$

$$H_1 : \text{at least one of the } \beta_k \text{ is non-zero}$$

Denoting the estimators of the model coefficients as $b_2, b_3, \ldots, b_K$, and constructing the following matrices:

$$_S = \begin{bmatrix} \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} \quad \boldsymbol{b}_S = \begin{bmatrix} b_2 \\ b_3 \\ \vdots \\ b_K \end{bmatrix}$$

And the variance-covariance matrix of estimated slope coefficients as:

$$\widehat{cov}(\boldsymbol{b}_S) = \begin{bmatrix} \widehat{var}(b_2) & \widehat{cov}(b_2, b_3) & \ldots & \widehat{cov}(b_2, b_K) \\ \widehat{cov}(b_3, b_2) & \widehat{var}(b_3) & \cdots & \widehat{cov}(b_3, b_K) \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{cov}(b_K, b_2) & \widehat{cov}(b_K, b_3) & \cdots & \widehat{var}(b_K) \end{bmatrix}$$

Then the hypotheses can be expressed in the compressed notation as:

$$H_0 : \boldsymbol{\beta}_S = 0$$

$$H_1 : \boldsymbol{\beta}_S \neq 0$$

If the null hypothesis is true, the outcome variable is not influenced by any of the predictor variables and the model reduces to:

$$y_i = \beta_1 + e_i$$

And, if the alternative hypothesis is true, then at least one of the predictor variables influences the outcome variable and should be included in the model.

It can be shown that the appropriate test statistic to test the hypotheses is an *F*-statistic expressed as:

$$F = \frac{(\boldsymbol{b}_S - \boldsymbol{\beta}_S)' \left[\widehat{cov}(\boldsymbol{b}_S)\right]^{-1} (\boldsymbol{b}_S - \boldsymbol{\beta}_S)}{K - 1} \sim F_{[(K-1),(N-K)]}$$

While testing for a single coefficient, the *t*-statistic can be thought of as the weighted measure of difference between the estimated and the true value of a parameter with the weight being the estimated variance $\widehat{var}(b_2)$, in simultaneous testing of all $K-1$ predictor variables, the *F*-statistic can be thought of as the weighted mean difference between the elements of

the estimator matrix $\boldsymbol{b}_S$ and those of the true coefficient matrix $\boldsymbol{\beta}_s$ with the weights being the estimated variance-covariance matrix $\widehat{cov}(\boldsymbol{b}_S)$.

To test the null hypothesis of all coefficients being nonzero, we compute, assuming the null hypothesis to be true, the value of the *F*-statistic as:

$$F = \frac{\boldsymbol{b}_S^{'}\left[\widehat{cov}(\boldsymbol{b}_S)\right]^{-1}\boldsymbol{b}_S}{K-1}$$

The decision rule for accepting or rejecting the null hypothesis are as usual: compare the estimated *F*-statistic value with the critical value for $(K-1),(N-K)$ degrees of freedom and if $F > F_{critical}$, we reject the null hypothesis.

Further, the *F*-statistic offers a slightly different perspective than the *R*-squared goodness of fit measure. While R^2 represents the ratio of explained variation to the total variation in the outcome variable accounted for the explanatory variable(s), it can be shown that the *F*-statistic represents the ratio of the explained variation to the unexplained variation in the outcome variable accounted for by the explanatory variables. Specifically:

$$F = \frac{\dfrac{\text{explained variation}}{(K-1)}}{\dfrac{\text{unexplained variation}}{(N-K)}}$$

The *F*-test for testing a null hypothesis comprising simultaneous multiple zero hypotheses can be thought of as a generalized case of testing a null hypothesis encasing a single zero hypothesis by the means of a *t*-test.

One may be tempted to assume that conducting one-to-one *t*-tests between all possible pairs of predictor variables is equivalent to conducting a single *F*-test—it is *not* so. While the argument and proof go beyond the scope of this book, one may keep in mind that it is quite possible for one-to-one *t*-tests to deduce some (or even all) estimated coefficients as statistically significant only for them to be nullified by the *F*-test.

Why use the *t*-test at all, then? It can be shown that $F = t^2$, and thus the *F*-statistic is agnostic about positive and negative values. This does not allow the *F*-test to perform a two-tailed test, while the *t*-test can do so. Many practical situations require one-tailed tests; only a *t*-test can do that.

3.1.7 Summary

A general linear econometric model having K unknown coefficients and $(K - 1)$ explanatory variables can be written as:

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \ldots + x_{iK}\beta_K + e_i$$

The coefficients of this model can be estimated using the ordinary least squares method. Reliable estimates are those close to true values with small standard errors, achieved by minimizing the sum of squared differences between observed and predicted values.

The average is taken as a basic predictive model and serves as a benchmark. Any other model must outperform it to be considered effective; the model whose predicted values are closer to the observed ones will be the better model. We compare $\left(Y - \overline{Y}\right)$ with $\left(Y - \hat{Y}\right)$. Thus, for a regression model: Looked at this way, the higher the proportion of variation explained by a model, the better it is compared to the *naïve* model. The proportion of variation explained by the model is:

$$\frac{ESS}{TSS} = \frac{\sum_1^N \left(\hat{Y} - \overline{Y}\right)^2}{\sum_1^N \left(Y_i - \overline{Y}\right)^2}$$

Where,

$$TSS = \sum_1^N \left(Y_i - \overline{Y}\right)^2 \left[Total\,Sum\,of\,Squared\,Errors\right]$$

$$ESS = \sum_1^N \left(\hat{Y} - \overline{Y}\right)^2 \left[Explained\,Sum\,of\,Squared\,Errors\right]$$

$$RSS = \sum_1^N \left(Y - \hat{Y}\right)^2 \left[Residual\,Sum\,of\,Squared\,Errors\right]$$

For a statistical model

$$TSS = ESS + RSS$$

$$\sum_1^N \left(Y_i - \overline{Y}\right)^2 = \sum_1^N \left(\hat{Y} - \overline{Y}\right)^2 + \sum_1^N \left(Y - \hat{Y}\right)^2$$

We define this ratio as the *coefficient of determination*:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

The values of R^2 lie between zero and one:

$$0 \le R^2 \le 1$$

The closer the value of R^2 to 1, the better the regression model estimates.

As the true population regression line is unknown, the parameter estimates of the sample regression line may be chance values, so we need to test the hypothesis that they are significantly non-zero:

$$H_0 : \beta_1 = \hat{\beta}_1$$

$$H_1 : \beta_1 \neq \beta_1$$

To test these hypotheses, we need to understand the characteristics of the estimator and its probability distribution. The mean of the ordinary least squares estimates of the sample regression line are:

$$E\left[\hat{\beta}_0\right] = \beta_0$$

$$E\left[\hat{\beta}_1\right] = \beta_1$$

The variance of the estimators can be given as:

$$var\left(\hat{\beta}_0\right) = \sigma^2\left[\frac{1}{N} + \frac{\bar{X}^2}{\Sigma x^2}\right] \; var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\Sigma x_i^2}$$

Also, for the error term, we have:

$$\boldsymbol{E}\left[\boldsymbol{e}\right] = 0$$

$$var\left(e_0\right) = \sigma^2\left[1 + \frac{1}{N} + \frac{x_0^2}{\Sigma x^2}\right]$$

$$cov\left(\boldsymbol{e}\right) = E\left[\boldsymbol{ee'}\right] = \sigma^2 \boldsymbol{I}_N$$

Thus, for a linear regression model:

$$Y \sim N\left(X\beta, \sigma^2 I_N\right)$$

$$e \sim N\left(0, \sigma^2 I_N\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{N} + \frac{\bar{X}^2}{\Sigma x^2}\right]\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\Sigma x_i^2}\right)$$

To check the statistical significance of the parameter estimates, the following hypotheses are constructed:

$$H_0 : \beta_0 = k$$

$$H_1 : \beta_0 \neq k$$

The test is carried out by calculating the *Z*-values of the estimates:

$$Z_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim N(0,1)$$

$$Z_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$$

and comparing them to the critical *Z*-value obtained for a particular level of significance (usually taken as $\alpha = 0.05$). in the *p*-value approach, if the $P\left(Z_{\hat{\beta}}\right) \leq P\left(Z_{critical}\right)$ we reject the null hypothesis and say that the estimates are statistically significant. In the confidence interval approach, we decide upon boundaries around the hypothesized mean for a particular level of confidence (given by $1-\alpha$):

$$\hat{\beta}_1 \pm Z_{critical} * \sigma_{\hat{\beta}_1}$$

For $\alpha = 0.05$, the 95% confidence interval can be constructed as:

$$\left(\hat{\beta}_1 - 1.96\sigma_{\hat{\beta}_1}\right) \leq \beta_1 \leq \left(\hat{\beta}_1 + 1.96\sigma_{\hat{\beta}_1}\right)$$

If the estimated value of the parameter lies within this boundary, the null hypothesis is accepted, else rejected. The *Z*-transformation is only possible if the population variance is known, else we do a *t*-transformation and the confidence interval is calculated as:

$$(\hat{\beta}_1 - t_c * \hat{\sigma}_{\hat{\delta}}) \leq \beta_1 \leq \left(\hat{\beta}_1 + t_c * \hat{\sigma}_{\hat{\delta}}\right)$$

The same decision rule applies for accepting or rejecting the null hypothesis.

To test the equality of parameters of two separate samples, we specify the complete model as:

$$Y_{Ai} = \beta_A + e_{Ai} \quad Y_{Bi} = \beta_B + e_{Bi}$$

$$e_{Ai} \sim N\left(0, \sigma^2\right) \quad e_{Bi} \sim N\left(0, \sigma^2\right)$$

$$cov\left(e_{Ai}, e_{Aj}\right) = 0 \quad cov\left(e_{Bi}, e_{Bj}\right) = 0 \quad i \neq j$$

$$cov\left(e_{Ai}, e_{Bj}\right) = 0$$

Then the ordinary least squares estimators for the two slopes are:

$$b_A = \frac{\Sigma_{i=1}^{N} Y_{Ai}}{N} \quad N\left(\hat{a}_A, \frac{\sigma^2}{N}\right)$$

$$b_B = \frac{\Sigma_{i=1}^{N} Y_{Bi}}{N} \quad N\left(\beta_B, \frac{\sigma^2}{N}\right)$$

The 'pooled' estimator for the common variance of the two populations is:

$$\hat{\sigma}^2 = \frac{\Sigma_{i=1}^{N}\left[\left(Y_{Ai} - b_A\right)^2 + \left(Y_{Bi} - b_B\right)^2\right]}{2\left(N-1\right)}$$

Then to test the hypotheses:

$$H_0 : \beta_A = \beta_B$$

$$H_1 : \beta_A \neq \beta_B$$

The appropriate test statistics is:

$$t_d = \frac{d}{\left(\frac{2\hat{\sigma}^2}{N}\right)^{\frac{1}{2}}} \sim t_{2(N-1)}$$

The null hypothesis being rejected if $|t| \geq t_{critical}$. The critical value is obtained from the table of *t*-distribution values for $2\left(N-1\right)$ degrees of freedom.

The *F*-statistic is the measure of overall fit of the linear regression model. To check the overall fit of the model, in other words, to check whether all predictor variables included in the model do indeed explain the variance in the outcome model, we hypothesize:

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_K = 0$$

$$H_1 : \text{at least one of the } \beta_k \text{ is non-zero}$$

To test the null hypothesis of all coefficients being nonzero, we compute, assuming the null hypothesis to be true, the value of the *F*-statistic as:

$$F = \frac{\boldsymbol{b}_S^{'}\left[cov\left(\boldsymbol{b}_S\right)\right]^{-1}\boldsymbol{b}_S}{K-1}$$

We compare the estimated *F*-statistic value with the critical value for $(K-1),(N-K)$ degrees of freedom and if $F > F_{critical}$, we reject the null hypothesis. *F*-statistic represents the ratio of the explained variation to the unexplained variation in the outcome variable accounted for by the explanatory variables:

$$F = \frac{\dfrac{\text{explained variation}}{(K-1)}}{\dfrac{\text{unexplained variation}}{(N-K)}}$$

The *F*-test can be thought of as a generalized case of a *t*-test.

3.1.8 Keywords

Theoretical Foundation: The construction of a regression model begins with the identification of outcome and predictor variables, followed by the postulation of a relationship between them. A theoretically sound model is essential for meaningful interpretation of results.

General Linear Econometric Model: This model represents the relationship between the outcome variable and predictor variables in a linear form. It can be represented in matrix form as: $y = X\beta + e$.

Ordinary Least Squares Estimator (BLUE): This estimator operates under specific assumptions and its reliability is judged based on theoretical justification, statistical tests of the estimates, and econometric considerations for the sample data.

Model Evaluation: The reliability of a regression model can be judged on three sets of criteria: a priori theoretical justification of the model, first order statistical tests of the estimates, and second order econometric considerations for the sample data on which the regression is run.

Total Sum of Squared Errors (TSS): It is the difference between the observed value of the outcome variable and its expected value estimated through the *average* model.

Explained Sum of Squared Errors (ESS): It is the difference between the estimated value of the outcome variable through the least squares method and the corresponding expected value estimated through the *average* model.

Residual Sum of Squared Errors (RSS): This is the difference between the observed value of the outcome variable and its estimated value through the least squares method.

Coefficient of Determination (R^2): This is the proportion of the variance in the dependent variable that is predictable from the independent variables.

Statistical Significance of Parameter Estimates: The parameter estimates might be chance values, so we need to test if they are significantly non-zero. This involves setting up null and alternate hypotheses for the slope of the sample regression line.

F-Statistic: The F-statistic is the measure of overall fit of the linear regression model. It represents the ratio of the explained variation to the unexplained variation in the outcome variable accounted for by the explanatory variables.

Pooled Estimator: The 'pooled' estimator for the common variance of the two populations is introduced in the document. It is used to test the equality of parameters of two separate samples.

3.1.9 Self-assessment Questions

1. Show that if the estimated confidence interval, $b \pm t_c \frac{\hat{\sigma}}{\sqrt{N}}$, contains β_0 within it, then the null hypothesis, $H_0 : \beta = \beta_0$, will not be rejected at the $(1-\alpha)$ level of confidence.
2. A coffee vending machine is calibrated to pour 7.0 ounces of Choco Latte into large coffee take-away cups. To check that the vending machine is operating as desired, an engineer pours four cups of Choco Latte and measures the volume of coffee poured and gets: 7.34, 6.92, 6.88 and 7.26 ounces. Assume that the coffee poured follows a normal distribution.
   a. Construct a 95% confidence interval around the average amount of coffee poured.
   b. Test the hypothesis, at the 5% level of significance, that the mean coffee poured is 7.0 ounces.
3. In the previous question (question number 2), following the initial check, the coffee vending machine is re-calibrated and again checked for mean pourings. A fresh pourings of Choco Latte into 5 cups yields the following volume of coffee poured: 7.93, 7.49, 7.33, 7.65, and 7.10 ounces. Assuming that the recalibration does not

affect the variance in the volume of Choco Latte poured, test the hypothesis that recalibration does not affect the mean volume of Choco Latte poured by the vending machine.

4. The Nike Shoe Company has launched a new pair branded 'Nike Airism' and wants to advertise that these shoes last at least 1000 kilometers of running. To test for this claim, they tested these shoes with six professional runners. The kilometers till their new shoes worn out were found to be: 1565, 1215, 850, 1020, 1170, and 1380 kilometers. Should the company make this claim?

a. Test the claim at 1% level of significance.

b. Construct a 95% confidence interval for the average number of kilometers that these shoes last.

5. Does studying sampling properties of an estimator make any sense when we typically draw only one sample for making inference?

6. Data for gross income and corporate tax (in billions of rupees) paid by the companies included in the BSE Sensex for two years, 2016-17 and 2017-18 is given in the table below:

2016-17		**2017-18**	
Income	**Tax**	**Income**	**Tax**
2.7762	1.0682	2.0731	1.3486
2.5386	1.4250	3.3184	1.2150
6.8234	0.8386	5.5981	2.1870
4.1461	0.3665	9.5096	2.1868
6.1798	1.7495	7.3467	1.2561
8.3197	1.6181	7.5085	2.0564
3.9525	1.7747	2.4499	2.1355
6.2480	1.8567	5.9901	2.0890
5.8963	1.8565	5.5270	1.8866
7.1567	0.5519	9.4755	2.0813
2.9713	1.7649	7.0346	0.9830
8.7398	0.4026	5.5228	1.6376

2016-17		2017-18	
Income	**Tax**	**Income**	**Tax**
7.1162	0.5971	2.1279	0.4925
7.1063	1.8568	9.3538	1.9689
7.4825	1.5432	10.3953	1.7745
7.0196	1.3294	8.9192	2.0775
4.7471	0.9791	10.0262	0.7272
3.1893	0.8428	4.1156	1.9481
4.7118	1.3834	9.8602	1.0169
9.1930	1.1728	6.8950	1.2389

a. Estimate the simple linear regression of tax over income for both the years separately.

b. Pool the observations from both the years and estimate the regression of tax over income for the pooled data as a whole.

c. Compare the estimates of the slope coefficients from (a) and (b).

7. Slope of the tax coefficient, in the question above (question number 6), represents the marginal tax rate. Test the hypothesis that the marginal tax rate for both the years is the same–at 5% level of significance.

8. As part of its continuous performance audit, a life insurance firm checks the performance of a randomly selected scheme from amongst its array of market-linked policies. It collects data from a group of 20 randomly selected policy holders and wants to model the relationship between family income and life insurance cover. The data collected (in lakhs of rupees per annum) is presented below:

Insurance	Income	Insurance	Income
386	40	112	83
332	40	391	49
273	94	110	74

137	42	254	64
225	25	100	46
263	59	367	73
107	27	288	78
193	28	151	41
329	67	296	48
376	33	347	95

a. Estimate a linear relationship between family income and insurance cover.

b. By how much does the insurance cover increase for every rise Rs.100,000 rise in family income?

c. At 5% level of significance, test whether the insurance cover coefficient is statistically significant.

9. While rice bran and soyabean both cost the same per kilogram, the oil produced by pressing them varies in volume. A vegetable oil manufacturing company wants to evaluate which of the two inputs would be cost effective for producing oil–assuming a fixed selling price that the customers are willing to pay. Following is the data of liters of oil produced per kilogram of inputs:

Input	**Output**	
	Soyabean	**Rice Bran**
1	1.40	0.82
2	3.25	1.32
3	3.19	3.16
4	2.20	4.56
5	1.06	2.21
6	2.63	0.98
7	3.67	2.66
8	4.19	3.42

Input	Output	
	Soyabean	Rice Bran
9	1.73	0.93
10	1.66	3.78
11	4.34	3.37
12	2.06	2.60
13	3.78	3.94
14	1.25	2.99
15	3.64	2.84

a. Estimate the marginal output for both soyabean and rice bran individually.

b. Test the hypothesis that both have the same mean liters of vegetable oil output.

c. Which input should the manufacturer use?

10. The following table presents the data on consumption (Y) and family income (X) for ten randomly selected family from a neighborhood:

Y	7	6	10	8	9	8	9	10	10	11
X	52	59	58	65	70	50	55	57	62	68

a. Estimate the consumption function by regressing consumption over family income.

b. What percentage of variance in the consumption expenditure is explained by the level of family income?

c. At 5% level of significance, test the hypothesis that the marginal propensity to consume is nonzero.

d. Construct a 95% confidence interval around the mean consumption level.

3.1.10 References

1. **Regression: Models, Methods and Applications** by Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, Brian D. Marx: This book provides an applied and unified introduction to parametric, nonparametric, and semiparametric regression. It bridges the gap between theory and application, featuring examples, applications, and user-friendly software.
2. **Statistics and Data Analysis for Financial Engineering** by David Rupert. This book provides theoretical results about linear least-squares estimation. The study of linear regression is facilitated by the use of matrices. It covers advanced topics in regression analysis.
3. **Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models** by Jim Frost. This book is one of the best regression books of all time. It provides a comprehensive guide to understanding, using, and interpreting linear models.
4. **Linear Models and Generalizations: Least Squares and Alternatives** by Rao et al. This book covers most of the topics with proofs and also covers generalized linear models (GLMs).
5. **The Coordinate-Free Approach to Linear Models** by Michael J. Wichura. This book provides a more geometric viewpoint to linear models. It may not cover all the topics but is a good resource for understanding the geometric aspects of linear models.

UNIT – IV: Assumptions of the Classical Linear Regression Model

Lesson 4.1 – Stochastic Assumptions of the Classical Linear Regression Model

Structure

4.1.1 The Classical Linear Regression Model

A linear regression model is a mathematical function expressing a relationship between, in its simplest form, one dependent and one or more independent variables of the form:

$$Y = f\left(X_1, X_2, \ldots, X_k\right)$$

where the relationship specifically takes the form of a straight line:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Giving this relationship a statistical form would mean the y and x–values would come from a sample and the coefficients of the model need to be *estimated* using this data. The coefficients, in this scenario, are *estimates* derived from a *sample* and may not necessarily be equal to their true values that *supposedly* exist in the *population*, breaks down the exactness of the relationship between y and x's. To mend for this very likely deviation, we introduce an unobserved variable, the error term, that will account for

any deviations between the actual (observed) and predicted (estimated) values. The statistical form of the relationship now becomes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{ki} X_k + \dot{o}_i$$

where, the subscript '*i*' represents the i^{th} observation in a series of '*n*' observations. Written in this form, the relationship is linear in independent variables; but in repeated sampling the X_i's are constant, and it is the coefficients, β's, that are estimated *variably*. This gives *linearity* another point of view: the relationship is linear in *parameters*.

Even this statistical relationship is devoid of any economic meaning; the estimated coefficients lack economic interpretation. For example, what does the derivative of the dependent variable with respect to an independent variable mean? Let's partially differentiate *y* with respect to, say, x_2:

$$\frac{\partial y}{\partial x_2} = \beta_2 + \frac{\partial \dot{o}_i}{\partial x_2}$$

One could possibly interpret this, had the partial derivative of the error term would have turned out to be zero. Had it been so, the partial derivative would have turned out to be:

$$\frac{\partial y}{\partial x_2} = \beta_2$$

But there is nothing in the model that would give:

$$\frac{\partial \dot{o}_i}{\partial x_2} = 0$$

But, even for β_2 above (and for that matter, any of the βs) to have an economic meaning, we need to make assumptions appropriate for the underlying data that is used to generate the estimates. The classical linear model makes a few assumptions that help ascribe some economic meaning to the model parameters, although most of those assumptions do not realistically apply to economic (or any other social science) data.

4.1.2 Fundamental Assumptions of the Model

The classical linear regression model makes two fundamental assumptions, both outlined above, but are worth restating. The first is that the model is a linear function of the parameters, thus, in matrix notation:

$$Y = \boldsymbol{X}\beta + \dot{c}$$

where Y is a vector of n observations of the outcome variable, $\boldsymbol{X}$ is a matrix of $n \times k$ predictors, β is a vector of $k+1$ parameters, and $\dot{c}$ is a vector of n error terms.

The second fundamental assumption relates to the way statistical significance of estimates is tested–repeated sampling to deduce distributional properties of the estimator. Thus, it assumes that in multiple sampling, the predictors are *nonstochastic*. In other words, $\boldsymbol{X}$ is a matrix of constants with rank k, and:

$$\lim \frac{1}{n} \boldsymbol{X}'\boldsymbol{X} = Q_X$$

To isolate effects of individual predictors, the Q_x is assumed to be a finite positive definite matrix. These assumptions prepare the ground for making the most crucial of all assumptions that sets the stage for not only devising an estimator for the coefficients but also impart them economic meaning: the error term is identically and independently distributed random variable with zero mean and a common variance. That is:

$$\dot{o} \sim IID\left(0, \sigma^2 I_n\right)$$

where $\dot{o}$ is a jointly distributed IID. Thus, the error terms have the same variance for every predictor variable (a property called homoscedasticity):

$$var\left(\dot{o}_i\right) = \sigma^2 \quad \forall i$$

Also, the error terms are not correlated with each-other:

$$covar\left(\dot{o}_i \dot{o}_j\right) = 0 \quad \forall i \neq j$$

An additional, but commonly invoked, assumption is that the error terms are, in fact, normally distributed, that is:

$$\dot{o} \sim N\left(0, \sigma^2 I_n\right)$$

Even when the underlying data is assumed to be IID, the assumption of normality does not, in general, hold for many economic and other social sciences data. For example, in a study where effect of petrol prices is being estimated, the dependent variable may be the number of times a service engineer goes out on field visits to address customer complaints for a consumer electronic company. If the petrol prices rise, the company may try to delay sending technicians to a locality from which only a

single complaint has been registered till a few more complaints are raised thus saving on the conveyance cost. This is a count variable, and such is discrete–not continuous–and thus cannot be normally distributed. Very frequently in Economics we have variables, such as number of units of a commodity consumed or a service delivered, that cannot assume negative values thus rendering them, technically, nonnormal. Nonetheless, this does not rule out assumption of normality as a good model under certain circumstances.

4.1.2.1 Randomness of the Error Term

Technically, error terms are necessitated by the fact that coefficient estimation takes place with sample data. The very use of a sample means that the outcome and predictors cannot be related with certainty–there always will be an element of uncertainty. There can be other sources of errors, too, such as: the model itself may be mis-specified, that is, the linear relationship may itself be not appropriate for modeling the relationship between the outcome and the predictors. This problem is accentuated in cases where theory does not guide about the mathematical structure of the relationship. Many of the real-life econometric phenomena are rife with this handicap.

Even if the relationship is correctly identified as being linear, the model constructed may miss some or many of the predictors—a problem of omitted variables. This omission may be due to sparse guidance from the theory, or because the concept of the predictor itself is nebulous such as the 'state of mind' of a consumer or even may be because the variables are difficult (if not impossible) to measure such as the 'complexity' of a design. This last problem translates into measurement errors wherein the predictor is not measured correctly either because it is abstract or because the scale of measurement is not proper. Errors in measurement can also occur if the measured values are not 'recorded' properly—this is a human error.

But the biggest source of error is the volatility of human nature which is the focal subject in most of the social science studies. The human nature has not yet been modeled with any certainty. Therefore, any measurement that involves human behavior–consumption pattern, work habits, investment decisions, political voting, etc.–is by its very nature, uncertain. Any attempt to model it has to deal with this uncertainty.

In face of such daunting circumstances, how to estimate the parameters of the model in any meaningful way? The way forward is if these errors are truly random then they can be accounted for in the estimation procedure and a reasonably reliable estimate be found. But the key lies in the error terms to be random. This is a difficult task and must be taken care of with utmost diligence. For the error terms to be random, the omitted variables must be unimportant (having minuscule or negligible effect on the outcome variable), must be numerous, and must all change in different directions for their overall effect on the outcome variable to be stochastic. This is a particularly important point because if the omitted variable is an important one, or they are only a handful in number or they all change more or less in the same way, then the errors will show a pattern and not be random. Moreover, the errors in recording the data must also not show any pattern, otherwise the omissions will be systematic and not random.

The discussion above should make it clear that there cannot be a formal statistical test for checking the randomness of the error terms. The true random errors (the ϵ's) are unobservable and they are estimated *assuming* randomness. This assumption is *endogenous* to the estimation the procedure(s) and as such the estimation procedure cannot itself be used to test for it.

4.1.2.2 Zero Mean of the Error Term

The assumption implies that the mean of the independently identically distributed error term is zero:

$$E(\grave{o}_i) = 0$$

The error term in any particular sample is thought of to be drawn from a distribution of error values that gets in repeated sampling. Let's consider a simple linear model with only one predictor:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \grave{o}_i$$

In repeated sampling, for a particular value of the predictor, say X_m, the error values will be different in different samples and these error values will follow a distribution with a mean of zero. Zero mean of the error term implies that:

$$E(Y) = \beta_0 + \beta_1 X_1 + E(\grave{o})$$

$$E(Y) = \beta_0 + \beta_1 X_1$$

Thus, the random part of the relationship drops off and the nonstochastic part can then be estimated using econometric procedures. This is reflected in the scatter of the Y and X values; if the values are scattered randomly *around* the population line, then only the sample regression line be a good approximation of the true relationship.

The effect of a nonzero random error term can be deduced as follows: suppose that the expected value of the error term is lesser than zero, that is,

$$E(\grave{o}_i) < 0$$

then the observed values of the outcome and the predictor from the sample drawn will lie *below* the true population regression line. Thus, the estimated regression line will be negatively *biased* and not reflect the true relationship.

Just as in the case of assumed randomness of the error term, its mean being zero cannot also be verified using any formal statistical test. This, again, happens because the estimation procedure sets the expected value of the error term to be zero at the very outset of deriving the estimator. And just as in the case of randomness, this assumption should be ensured at the very beginning of the model building and data collection efforts. The researcher should ensure that no important predictor is omitted, scales of measurement are appropriate, and the recording errors, if any, do not follow any pattern. Eliminating avenues of systematic variation is the only way to ensure that the assumption of zero mean of the random error holds.

Despite all efforts, in many cases, while working with real-life data a researcher is obliged to omit variables that may not be unimportant–this happens particularly when the data exhibits multicollinearity (we will discuss it in a later lesson). Also, many of the economic variables tend to vary systematically, e.g., gross domestic product and corporate tax revenue.

4.1.2.3 Equal Variance of the Error Term

The assumption of equal variance goes a step ahead of the assumption of zero mean and implies that across all values of the predictors, the probability distribution of the error term has the same–common–variance. Thus, as we have discussed earlier, in repeated sampling the error term for a particular predictor in a particular sample is thought of as coming from a probability distribution of the error term that has a mean zero and this is true for all predictors in the model. The equal variance assumption states

that these probability distributions of error terms for each of the distinct predictors have a constant, equal variance. That is:

$$var(\grave{o}_i|X_i) = \sigma^2 \quad \forall i$$

Again, this means that for a multiple regression model of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \grave{o}_i$$

the variance of the distribution of error term associated with, say X_2 is the same as the variance of the distribution of error associated with X_j or of any other predictor; the variance of the distribution of error term does not change with changing the predictor.

The equal variance property is called homoscedasticity and can be graphically represented as in the following figure:.

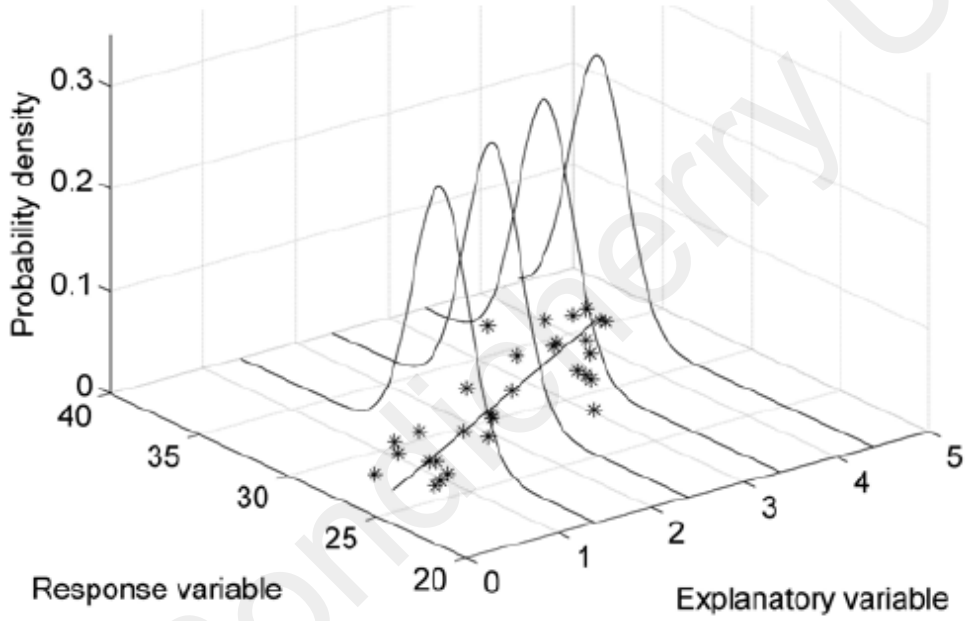


Figure 1: Homoscedastic Distribution of Errors

This figure shows that the spread of the residuals around the regression line is consistent, regardless of the independent variable's values.

Under homoscedasticity, the Ordinary Least Squares (OLS) estimator for the regression coefficients (β) is efficient. This means it has the minimum variance among all unbiased estimators—an integral part of the OLS estimator's property of being BLUE.

Homoscedasticity allows for more precise confidence intervals and hypothesis tests. When the error terms have constant variance, the calculated standard errors of the regression coefficients are reliable.

These standard errors are used to construct confidence intervals and test hypotheses about the β coefficients. Inaccurate variance estimates lead to unreliable standard errors, hindering proper inference.

A battery of tests exists for confirming homoscedasticity in the dataset. If the dataset has unequal variances for the error terms, that is, if:

$$var\ (\in_i | X_i) = \sigma_i^2$$

then the variance of the distribution of errors is changes with a change of X; it is not independent of the predictors. This condition is called heteroscedasticity. Detecting heteroscedasticity is crucial to ensure reliable results and guide appropriate remedial measures. Some of the most commonly used tests for confirming homoscedasticity (or equivalently, detecting heteroscedasticity) are:

Spearman's Rank Correlation Test

The simplest (computationally) and most widely applicable (can be performed with any sample size, large or small) of all tests. It computes the rank correlation of estimated errors, with high rank correlation suggesting presence of heteroscedasticity. The procedure for conducting the rank correlation test it is as follows:

1. Regress the outcome variable, Y, on the predictor, X:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \text{ò}_i$$

   and obtain the error estimates, e's, of the model residuals, ċ's.
2. Take the absolute values of the e's and arrange them in ascending order (or descending order) and assign ranks to the ordered values. Repeat the procedure for the predictor values, X's.
3. Compute pair-wise difference in ranks (D_i) for all pairs of (X_i, e_i) in the sample of n observations.
4. Compute the rank correlation according to the Spearman's formula:

$$\rho_{X,e} = 1 - \frac{6\Sigma D_i^2}{n\left(n^2 - 1\right)}$$

For a model with more than one predictor, the rank correlation needs to be calculated for each pair of the estimated errors and predictors, separately. Then an F-test can be done to check the hypothesis that the pair-wise rank correlations are all simultaneously nonzero.

$$H_0 : \rho_{X_1,e_1} = \rho_{X_2,e_2} = \ldots = \rho_{X_k,e_k} = 0$$

H_1 : at least one paire-wise rank correlation is nonzero

A statistically significant high rank correlation signifies presence of heteroscedasticity in the dataset. Spearman's rank correlation is used instead of Pearson's coefficient of correlation because in running the ordinary least squares procedure for estimating the residuals, we assume:

$$\Sigma eX = 0$$

therefore, the Pearson's coefficient of correlation computed using the formula:

$$r_{X,e} = \frac{\Sigma Xe}{\sqrt{\Sigma e^2}\sqrt{\Sigma x^2}}$$

will always yield $r_{X,e} = 0$ since the numerator in the formula is zero.

Glejser Test

One solution to $\Sigma eX = 0$ is to take absolute values of the residuals, $|e_i|$. Also, in modeling regression, it the dispersal of the values around the regression line is all that matters, and not its direction (above or below the regression line). This idea is used by the Glejser test. But instead of finding the coefficient of correlation, the test runs a regression of the absolute vales of the error terms on the associated predictors. But there are no *a priori* grounds for modeling a specific form of the regression relationship between the residuals and the predictors, different versions bearing different exponentiations of the predictor variables are experimented with to arrive at a satisfactory model:

$$|e| = c_0 + c_1 X_j^2$$

$$|e| = c_0 + \frac{c_1}{X_j}$$

$$|e| = c_0 + c_1\sqrt{X_j}$$

Once a satisfactory model is decided upon, the regression is run for each pair of absolute errors and predictors. The values of the coefficients (c_0, c_1) thus obtained are checked for their statistical significance, either by conducting individual *t*-tests or a comprehensive *F*-test. It is only in the case that both the coefficients turn out to be not significantly different from zero (i.e., $c_0 = c_1 = 0$), can we conclude data being homoscedastic. Otherwise, if

both $c_0 \neq 0$ and $c \neq$, it is the case of *mixed* heteroscedasticity. And if $c_0 = 0$ but $c_1 \neq 0$, we call it *pure* heteroscedasticity.

The Glejser's test has an added advantage, while checking for heteroscedasticity, we already check for the exact form of heteroscedasticity, i.e., the way in which the error terms are related with the predictors. This is an important aspect as it shapes the way in which we possibly could attempt to correct for estimation of regression parameters in the presence of heteroscedasticity.

Breusch-Pagan Test

The Breusch-Pagan test is predicated on the premise that if heteroscedasticity is present, the variance of the errors from a regression model will be correlated with one or more of the independent variables. The test involves regressing the squared residuals on the independent variables and the predicted values of the dependent variable, and testing the null hypothesis that the coefficients on the predicted values are simultaneously equal to zero.

The Breusch-Pagan test is a Lagrange multiplier (LM) test that utilizes the squared residuals from the initial regression to assess the presence of heteroscedasticity. It tests the null hypothesis of homoscedasticity against the alternative of heteroscedasticity. The Breusch-Pagan test statistic (LM statistic) is calculated as follows:

$$LM = nR^2$$

$$= \sum \frac{e_i^2}{\sigma^2}$$

where:

$n =$ number of observations

$R^2 =$ coefficient of determination from the initial regression

$e_i =$ residuals from the initial regression

$\sigma^2 =$ assumed constant variance of the error term.

The procedure for conducting the test is as follows:

1. Regress the dependent variable *Y* on the independent variables *X* to obtain the residuals, e.
2. Compute the squared residuals e^2.

3. Regress e^2 on the original independent variables X and obtain the R-squared value, R^2.
4. Calculate the test statistic LM = nR^2. This statistic follows a chi-square (χ^2) distribution with degrees of freedom equal to the number of independent variables.
5. Compare *LM* to the critical value from the χ^2-distribution to determine the presence of heteroscedasticity.

Large *LM* statistics with a *p*-value less than a chosen significance level (e.g., $\alpha = 0.05$) indicate a rejection of the null hypothesis and suggest heteroscedasticity.

Goldfeld-Quandt Test

The Goldfeld-Quandt test is a test for heteroscedasticity that is based on the idea that if the error term is heteroscedastic, then the variance of the error term will be different for different subsets of the data. The test involves dividing the data into two or more subsets based on the values of the independent variables and comparing the variance of the error term in each subset. The null hypothesis is that the variance of the error term is the same in all subsets, indicating homoscedasticity. The alternative hypothesis is that the variance of the error term is different in at least two subsets, indicating heteroscedasticity. The procedure for running the test is as follows:

1. Sort the data (comprising *k* variables) by the independent variable suspected of influencing the heteroscedasticity.
2. Divide the sorted data into two groups, excluding a central portion (*c*, *n*–*c*) to ensure no overlap.
3. Perform separate OLS regressions for each group and compute the sum of squared residuals (*SSR*) for both.
4. Calculate the test statistic as the ratio of the larger *SSR* to the smaller *SSR*, which follows an *F*-distribution.

$$F^* = \frac{\frac{\Sigma e_2^2}{\frac{n-c}{2}-k}}{\frac{\Sigma e_1^2}{\frac{n-c}{2}-k}} = \frac{\Sigma e_2^2}{\Sigma e_1^2}$$

The degrees of freedom being: $\frac{n-c}{2}-k$, same for both the numerator and denominator.

5. Use the *F*-distribution to assess the significance of the test statistic. If $F^* > F_{critical}$, error terms are heteroscedastic, else if $F^* < F_{critical}$, they are homoscedastic. If the distribution of error terms is homoscedastic, the F^* value will be closer to 1; the more heteroscedastic, the higher the F^* value.

Goldfeld-Quandt test can only be performed on large sample data. The minimum viable size of the sample is defined as the one where the number of cases observed (n) is at least twice as many as the number of regression coefficients (β's) to be estimated. Another limitation of this test is that it assumes away one of the major and commonly observed problems with social sciences data—presence of serial correlation among the error terms.

White's Test

White's test is a general test for heteroscedasticity that can detect any form of non-constant variance of the error term. It is based on the idea that if the error term is heteroscedastic, then the squared residuals will be correlated with the independent variables. The test involves regressing the squared residuals on the independent variables, their squared terms, and their cross-product terms. The null hypothesis is that the coefficients on all these terms are simultaneously equal to zero, indicating homoscedasticity. The alternative hypothesis is that at least one of these coefficients is non-zero, indicating heteroscedasticity.

Also known as the heteroscedasticity-consistent (HC) test, is an *LM* test that utilizes the squared residuals along with the product of the independent variables (X_iX_j) to detect heteroscedasticity. It is more general than the Breusch-Pagan test and does not require specifying a particular form of heteroscedasticity. This test is robust to the presence of non-normal errors. While the test requires additional computations compared to the Breusch-Pagan test, its robustness to non-normal errors can be advantageous in certain situations.

Park Test

The Park test assesses the relationship between the variance of the errors and one of the independent variables in a logarithmic form. This test

requires regressing Y_i's on X_i's to obtain residuals, e, and then taking the natural logarithm of the absolute residuals and regressing it on the natural logarithm of one of the independent variables suspected of influencing the variance. A significant relationship indicates heteroscedasticity.

Graphical Assessment: Before conducting formal tests, it is often valuable to visually assess the presence of heteroscedasticity. Plot the residuals against the predicted values (fitted values) to observe any patterns suggesting unequal variance. A fan-shaped or cone-shaped pattern in the residual plot is a common indicator of heteroscedasticity. The appropriate heteroscedasticity test depends on the specific assumptions one is willing to make about the nature of the heteroscedasticity.

- Breusch-Pagan test is preferred when assuming a linear relationship between the error variance and the independent variables.
- White's test is more robust when the exact form of the heteroscedasticity is unknown.
- Park's test is helpful in detecting heteroscedasticity arising from omitted lagged dependent variables in the model.

These tests are based on different ideas and assumptions, and they can detect different forms of heteroscedasticity. It is important to choose the appropriate test based on the nature of the data and the research question, and to interpret the results carefully.

4.1.2.3.1 Consequences of Heteroscedasticity

The estimates of the parameters are not affected by heteroscedasticity as unbiasedness of an estimator is not dependent on constant variance of the error term. All that is required is that:

$$cov\left(x_i \dot{o}_i\right) = 0 \forall i$$

Since we have already shown that:

$$b_1 = \beta_1 + \frac{\Sigma x_i \dot{o}_i}{\Sigma x_i^2}$$

Therefore, the expected value of the parameter estimate, b_1, will give:

$$E\left(b_1\right) = \beta_1 + E\left[\frac{\Sigma x_i \dot{o}_i}{\Sigma x_i^2}\right] = \beta_1$$

Also, for the intercept, we have:

$$b_0 = \bar{Y} - b_1\bar{X}$$
$$= \left(\beta_0 + \beta_1\bar{X} + \bar{\varepsilon}\right) - b_1\bar{X}$$

Taking expected values, we get:

$$E\left(b_0\right) = \beta_0 + \beta_1\bar{X} + E\left(\bar{\varepsilon}\right) - E\left(b_1\bar{X}\right) = \beta_0$$

Despite retaining its unbiasedness, the ordinary least squares estimator in the presence of heteroscedasticity, no longer has the least variance among the set of all unbiased estimators. That is, it is no longer *best*. This, in turn, makes the predicted value of the outcome variable inefficient. In other words, for a given value of the explanatory variable, the predicted value of the outcome variable will have high variance as it would subsume variances of the true residuals as well as of the parameters themselves.

One consequence of this is that we can no longer test the statistical significance of the parameter estimates or construct confidence intervals around them. We know that variances of the parameter estimates are given as:

$$var\left(b_0\right) = \sigma^2 \frac{\Sigma X^2}{n\Sigma x^2}$$
$$var\left(b_1\right) = \sigma^2 \frac{1}{\Sigma x^2}$$

In both the equations above, the variance of the residuals could be taken out since it was constant; in case of heteroscedasticity, it is no longer so. The variance of the estimates will change with change in the values of the explanatory variable(s). Therefore, the formulas to calculate variances of the estimates are no longer applicable in the presence of heteroscedasticity and we cannot test the hypotheses of the parameter estimates to be nonzero.

4.1.2.3.2 Solutions for Heteroscedastic Data

A few commonly used methods researchers use to deal with heteroscedasticity are:

Transforming the Dependent Variable: In some cases, heteroscedasticity can be stabilized by applying a transformation to the dependent variable, such as the logarithmic, square root, or inverse transformation. For example:

- Log transformation: $Y^* = \log\left(Y\right)$

- Square root transformation: $Y^* = \sqrt{Y}$
- Inverse transformation: $Y = \frac{}{Y}$

This approach can help reduce heteroscedasticity, but it may also introduce interpretation challenges and complicate the model. Additionally, they may not work for zero or negative values of the dependent variable without adjustments.

Weighted Least Squares (WLS): WLS provides an alternative to Ordinary Least Squares (OLS) by weighting each observation inversely proportional to its variance. This method involves transforming the original regression model by dividing each observation by the square root of the corresponding variance of the error term.

The rationale behind WLS is to give more weight to observations with smaller variances and less weight to observations with larger variances. Let us define the original linear regression model as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where y_i is the dependent variable, x_i is the independent variable, β_0 and β_1 are the regression coefficients, and ε_i is the error term with $var(\varepsilon_i) = \sigma_i^2$. The WLS estimator minimizes the following weighted sum of squared residuals:

$$\frac{\Sigma_i \left(y_i - \beta_0 - \beta_1 x_i\right)^2}{\sigma_i^2}$$

The WLS estimators of β_0 and β_1 are given by:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{WLS}} = \left(\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\Omega}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{\Omega}^{-1} \boldsymbol{y}$$

where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{y}$ is the vector of dependent variables, and $\boldsymbol{\Omega}$ is the diagonal matrix with σ_i^2 as the diagonal elements. The limitation of WLS is that estimating the correct weights can be challenging and might require prior knowledge or assumptions about the variance structure.

Heteroscedasticity-Consistent Standard Errors (HCSE): HCSE are robust standard errors that adjust for heteroscedasticity without needing to specify a particular form of heteroscedasticity, allowing for valid hypothesis testing. This approach involves estimating the regression coefficients using Ordinary Least Squares (OLS) but correcting the standard errors of the estimates to account for heteroscedasticity.

The rationale is to obtain valid statistical inferences without transforming the original model. Several variants of HCSE exist, such as HC0, HC1, HC2, and HC3, which differ in their assumptions about the error distribution and the way they estimate the $\boldsymbol{\Omega}$ matrix. The basic HC0 estimator adjusts the OLS variance-covariance matrix as follows:

$$\boldsymbol{var}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}\right)=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\Omega}\boldsymbol{X}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}$$

where $\boldsymbol{\Omega}$ is the diagonal matrix with ε_i^2 as the diagonal elements, and ε_i are the OLS residuals.

While HCSE estimators make standard errors more reliable under heteroscedasticity, they do not address the inefficiency of the coefficient estimates themselves.

Generalized Least Squares (GLS): GLS extends OLS by assuming a specific form of heteroscedasticity and possibly correlation among the error terms, leading to more efficient estimates when the form of heteroscedasticity is correctly specified.

This method involves transforming the original regression model to obtain a new model with homoscedastic errors, and then applying OLS to the transformed model. The rationale is to directly address the heteroscedasticity issue and obtain efficient parameter estimates. To illustrate, let's assume that the variance of the error term is a known function of the independent variables:

$$var\left(\varepsilon_i\right)=\sigma^2 g\left(x_i,\theta\right)$$

where $g\left(x_i,\theta\right)$ is a specified function of the independent variables x_i and a vector of unknown parameters θ. The GLS estimator involves dividing the original model by the square root of $g\left(x_i,\theta\right)$:

$$\frac{y_i}{\sqrt{g\left(x_i,\theta\right)}}=\frac{\beta_0}{\sqrt{g\left(x_i,\theta\right)}}+\beta_1\left(\frac{x_i}{\sqrt{g\left(x_i,\theta\right)}}\right)+\frac{\varepsilon_i}{\sqrt{g\left(x_i,\theta\right)}}$$

The GLS estimator of β_0 and β_1 is then obtained by applying OLS to the transformed model. The limitation of GLS is that it requires knowledge or assumptions about the form of the variance-covariance matrix, which might not be accurately known in practice.

Robust Standard Errors: This approach involves estimating the regression coefficients using OLS and then computing standard errors

that are robust to heteroscedasticity and potential misspecification of the error distribution. The rationale is to obtain valid statistical inferences without making strong assumptions about the error distribution or the form of heteroscedasticity. The robust standard errors are computed using the Huber-White sandwich estimator:

$$var\left(\hat{\beta}_{OLS}\right) = \left(X^T X\right)^{-1} X^T \Psi X \left(X^T X\right)^{-1}$$

where Ψ is a diagonal matrix with ψ_i as the diagonal elements, and ψ_i are the squared residuals multiplied by a weight function that downweighs large residuals.

The limitation of robust standard errors is that they do not improve the efficiency of the OLS estimator; they only provide valid statistical inferences under heteroscedasticity and potential misspecification of the error distribution.

Bootstrapping: Bootstrapping is a non-parametric approach that involves repeatedly resampling the dataset with replacement and estimating the model on each sample. It can provide robust estimates of the standard errors without assuming a particular form of heteroscedasticity.

Bootstrapping can be computationally intensive, especially for large datasets or complex models. Moreover, its effectiveness depends on the representativeness of the bootstrap samples and the stability of the model estimates across these samples.

Each method has its strengths and limitations, and the choice among them depends on the specific circumstances of the data and research question. In practice, researchers often start with diagnostic tests to detect the presence of heteroscedasticity and then choose an appropriate method based on the nature of the problem and the available information about the error variance structure. Often, researchers may employ more than one of these methods to robustly check the consistency of their findings under different assumptions about the variance of the residuals.

When dealing with heteroscedasticity, it is crucial to diagnose the issue correctly and choose the appropriate remedy based on the specific characteristics of the data and research question. It is also essential to ensure that the assumptions of the chosen method are met to obtain valid and reliable results.

4.1.2.4 Normally Distributed Error Term

The classical statistical tests used to perform checks on the estimated coefficients of the regression model are all parametric in nature, meaning thereby that they assume certain form of probability distribution for the data they analyze. Most parametric tests assume that the test statistics that they calculate are distributed normally or the specific distribution that the statistics follow are derived from some form of normal distribution. For example, the χ^2 distribution arises from the sum of the squares of independent standard normal random variables; the *F*-distribution is generated by a ratio of two random variables that follow χ^2 distribution.

Recall that we found that the ordinary least squares estimator is distributed normally with its expected value (mean) being the true population parameter:

$$b \sim N\left(\beta, \sigma^2 \frac{\Sigma X^2}{n\Sigma x^2}\right)$$

where σ^2 is really the variance of the residuals, which in turn is assumed to be distributed normally with its mean equal to zero:

$$\text{ò}_i \sim N\left(0, \sigma^2\right) \forall i$$

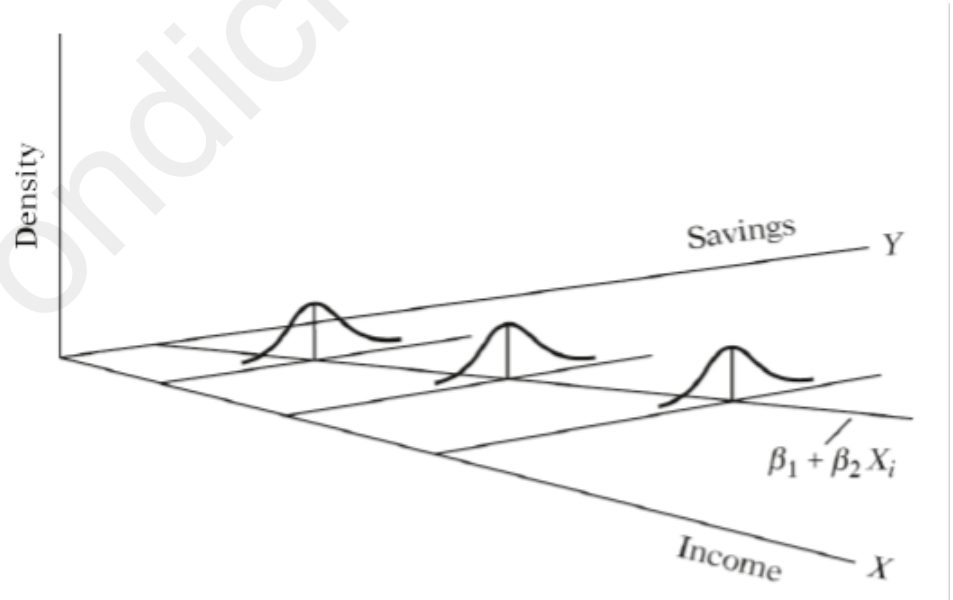


Figure 5: Normally Distributed Errors

Therefore, our ability to test statistical significance of regression estimates stems from the error terms being distributed normally. Thus, irrespective of the *X* values, the distribution of ò's not only has a common mean (0)

and common variance (σ^2), but is also distributed normally:

There are indirect ways of checking this; indirect because the error terms are not observable. Following are the most commonly employed tests to assess the normality of residuals:

Shapiro-Wilk Test

This test is a popular choice for assessing the normality of residuals. It calculates a *W*-statistic based on the ordered residuals and compares it to a critical value from a reference table.

$$W = \frac{\left(\Sigma_{i=1}^{n} a_i x_{(i)}\right)^2}{\left(\Sigma_{i=1}^{n} a_i - \overline{x}_i\right)^2}$$

Where, $x_{(i)}$ is the i^{th} order statistic and the coefficients, a_i's are computed as:

$$\left(a_1, a_2, \ldots, a_n\right) = \frac{m^T V^{-1}}{C}$$

Where:

$$C = \left(m^T V^{-1} V^{-1} m\right)^{\frac{1}{2}}$$

$$m = \left(m_1, m_2, \ldots, m_n\right)^T$$

The null hypothesis is that the residuals are normally distributed. The test performs well for smaller sample sizes (n < 50) and the result is relatively easy to interpret and implement. But the major drawback of this test is that it is less powerful for larger sample sizes, where even minor deviations from normality might be flagged as significant and its results are sensitive to outliers in the residuals.

Kolmogorov-Smirnov Test (K-S Test)

This non-parametric test compares the empirical cumulative distribution function (ECDF) of the residuals with the theoretical CDF of a normal distribution. It calculates a D-statistic representing the maximum absolute difference between the two distributions.

$$D_n = \sup_x \left|F_n\left(x\right) - F\left(x\right)\right|$$

Where $\sup_x$ is the *supremum* of the set of distances and $F_n\left(x\right)$ is defined as an IID function:

$$F_n(x) = \frac{1}{n}\sum_{i=1} 1_{(-\infty,x]}(X_i)$$

The test makes no assumptions about the underlying distribution and can be used for small sample sizes. However, it is less powerful than the Shapiro-Wilk test for normality. The results are also sensitive to the choice of bin size in the ECDF calculation.

Anderson-Darling Test

This test utilizes a squared difference between the ECDF of the residuals and the theoretical normal CDF. It provides a statistic (A^2) that is asymptotically chi-square distributed under the null hypothesis of normality.

$$A^2 = -n - S$$

Where:

$$S = \sum_{i=1}^{n} \frac{2i-1}{n}\left[ln\left(F\left(Y_i\right)\right) + ln\left(1 - F\left(Y_{n+1-i}\right)\right)\right]$$

The nonparametric counterpart, k-samples test, computes the test statistic as:

$$A_{kN}^2 = \frac{1}{N}\sum_{i=1}^{k}\frac{1}{n_i}\sum_{j=1}^{N-1}\frac{\left(NM_{ij} - jn_i\right)^2}{j\left(N-j\right)}$$

It is more powerful than the K-S test for normality assessment, its major advantage being sensitive to deviations from normality across the entire distribution, not just the tails. But this sensitivity makes it more sensitive to sample size, too, than the Shapiro-Wilk test—it requires larger sample sizes for reliable results.

Normal Quantile-Quantile Plot

This graphical technique is a valuable tool for visually assessing the normality of residuals. It plots the quantiles of the residuals against the quantiles of a standard normal distribution. Ideally, the points should fall roughly along a straight line, indicating a normal distribution. Deviations from the line suggest non-normality (see Figure–3).

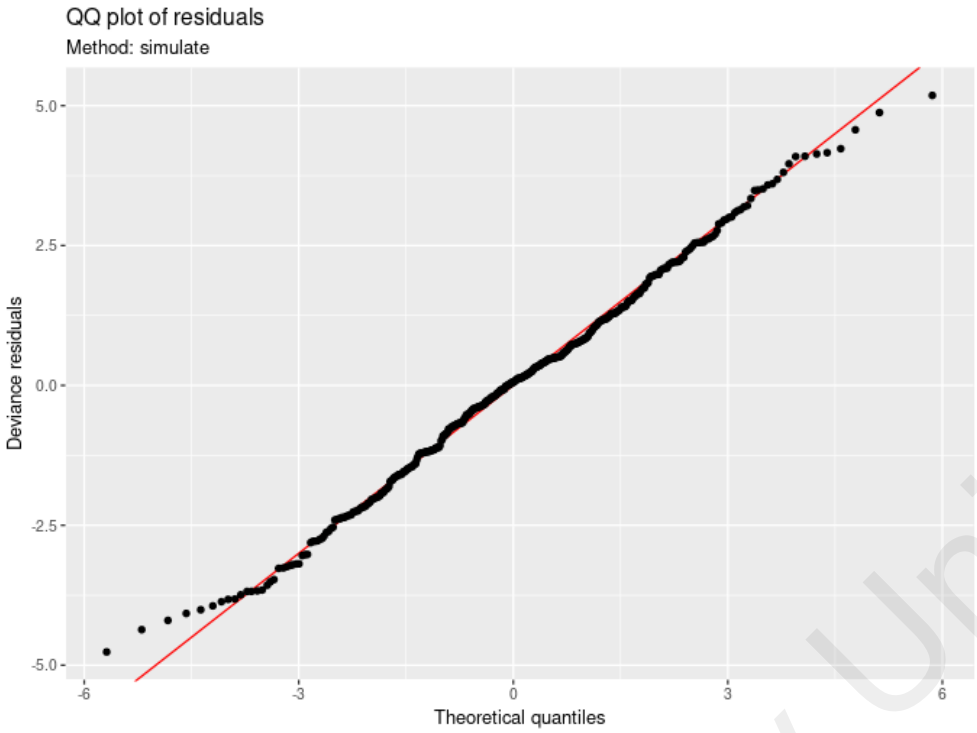


Figure 6: Normal Q-Q Plot of the Residual

This Photo by Unknown Author is licensed under CC BY

The appropriate test for normality of residuals depends on various factors, including:

- **Sample Size:** The Shapiro-Wilk test is preferred for smaller samples, while the Anderson-Darling test might be better for larger samples.
- **Sensitivity to Outliers:** The Shapiro-Wilk test is more sensitive to outliers than the K-S test.
- **Visual confirmation:** Regardless of the chosen test, a normal Q-Q plot is highly recommended for visual confirmation of the normality assumption.

While these tests assess normality, a slight deviation might not be a major concern, especially for larger sample sizes. The focus should be on substantial departures from normality that could invalidate inferences. Transformations of the data (e.g., log transformation) can sometimes achieve normality if the non-normality arises from a specific functional form.

4.1.3 Summary

The Classical Linear Regression Model (CLRM) forms the foundation of econometric analysis, providing a structured approach to

understanding relationships between a dependent variable and one or more independent variables. At its core, CLRM represents a mathematical equation that links a dependent variable, Y, representing the outcome of interest, to independent variables, $X_1, X_2, \ldots, X_k$, which are thought to influence Y. The model is linear, both in parameters ($\beta_0, \beta_1, \ldots, \beta_k$) and variables, simplifying its estimation and interpretation. The equation, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + ċ$, captures this relationship, where ċ signifies the error term accounting for deviations from the estimated model due to factors not included in the regression.

Central to the utility and validity of CLRM are its assumptions, which guarantee the model's estimators are the Best Linear Unbiased Estimators (BLUE). These assumptions include the linearity in parameters, the error term's independence, and identical distribution with a zero mean, $IID\left(0, \sigma^2\right)$, homoscedasticity (constant variance of error terms), and no perfect multicollinearity among independent variables.

A crucial feature of CLRM is the error term \epsilon, which embodies all the variation in *Y* not explained by the model. This includes omitted variable bias, measurement errors, and the intrinsic randomness of the dependent variable. For the model to provide unbiased and efficient estimates, it's imperative that the error term is IID with a zero mean and constant variance. This ensures that on average, the model does not systematically overestimate or underestimate the dependent variable, and the precision of estimates remains consistent across observations.

The assumption of no perfect multicollinearity is vital for the estimation process. It ensures that each independent variable contributes unique information to explaining *Y*, allowing for the clear identification and estimation of their effects. Violating this assumption would render some parameters indeterminate, as the linear relationship among some independent variables could infinitely satisfy the regression equation.

Moreover, the assumption that the error term is normally distributed facilitates the application of various statistical tests, such as t-tests and F-tests, which rely on normality assumptions for deriving their distributions under the null hypothesis. Although this assumption is not strictly necessary for the OLS estimator to be BLUE, it is crucial for conducting hypothesis testing and constructing confidence intervals around the parameter estimates.

In practice, these assumptions are tested using diagnostic tests and graphical analysis. For instance, the presence of heteroscedasticity can be examined through plots of residuals versus fitted values or using formal tests like the Breusch-Pagan or White test. Similarly, the normality of residuals can be assessed using the Shapiro-Wilk or Kolmogorov-Smirnov tests, among others. When assumptions are violated, alternative estimation techniques or transformations of the data may be employed to rectify the issues, ensuring the robustness of the model's estimates.

While CLRM provides a powerful tool for analyzing linear relationships, careful attention must be paid to its underlying assumptions. Adherence to these assumptions ensures that the model yields reliable, interpretable, and meaningful results that can inform decision-making and contribute to our understanding of complex phenomena in economics and other disciplines. Through empirical testing and adjustment, researchers can address the limitations of the classical model, enhancing the accuracy and applicability of their analyses in real-world scenarios.

4.1.4 Keywords

Classical Linear Regression Model (CLRM): A statistical framework for modeling the relationship between a dependent variable and one or more independent variables in a linear fashion.

Nonstochastic Independent Variables: The idea that independent variables are fixed in repeated sampling, not random.

Randomness of the Error Term: The principle that the errors in the predictions of the dependent variable are random and not systematic.

Zero Mean of the Error Term: The assumption that the average of the error terms is zero, ensuring unbiased estimates.

Homoscedasticity: The assumption that the error term has a constant variance across all levels of independent variables.

Heteroscedasticity: A violation of the equal variance assumption, where the error term's variance changes across the range of independent variables.

Normally Distributed Error Term: The assumption that the error term is normally distributed, facilitating certain statistical tests.

IID (Independent and Identically Distributed): Describes the error terms having the same probability distribution and being mutually independent.

Autocorrelation (Serial Correlation): A condition where error terms are correlated with each other across observations, violating the assumption of independent errors.

Diagnostic Testing: Procedures used to check whether the assumptions of the CLRM are met in the estimated model, including tests for homoscedasticity, normality of errors, and autocorrelation.

Spearman's Rank Correlation Test: A nonparametric test used to detect heteroscedasticity by assessing the correlation between the ranks of absolute residuals and the ranks of predictor variables.

Breusch-Pagan Test: A test for heteroscedasticity that regresses squared residuals on independent variables and their squared values to check if variances of the errors are constant.

White's Test: A test for heteroscedasticity that does not require specifying a model for the variance of the errors. It uses the squared residuals from the regression model as dependent variables in a new regression against the original independent variables and their squares and cross-products.

Durbin-Watson Test: A statistical test used to detect the presence of autocorrelation (specifically, first-order serial correlation) in the residuals from a regression analysis.

Glejser Test: A test for heteroscedasticity that involves regressing the absolute values of the residuals from the original regression on the independent variables or functions of them to check for a systematic relationship between the variance of the errors and the independent variables.

Goldfeld-Quandt Test: A test for detecting heteroscedasticity by dividing the data set into two or more groups and comparing the variances of the errors across these groups.

Park Test: A test for heteroscedasticity that involves regressing the log of the squared residuals on the log of one or more independent variables to identify a variance that changes with the level of an independent variable.

Lagrange Multiplier (LM) Test: A general test used in econometrics to test for the presence of omitted variables, autocorrelation, heteroscedasticity, and other misspecifications in a regression model.

4.1.5 Self-assessment Questions

1. What does the assumption of linearity in parameters mean for a CLRM?
2. Explain why the assumption of the error term having zero mean is crucial in CLRM.
3. Describe homoscedasticity. Why is it important in regression analysis?
4. Define the term "Best Linear Unbiased Estimators" (BLUE). What conditions must be met for OLS estimators to be BLUE?
5. What are the implications of violating the normality assumption of the error term in CLRM?
6. What statistical test can be used to detect the presence of heteroscedasticity in a regression model?
7. Why is it problematic if independent variables are not nonstochastic in the context of CLRM?
8. Given a simple regression output where the estimated regression equation is $Y=2+3X$, calculate the predicted value of Y when $X=4$.
9. A regression model reports $R^2=0.85$. Explain what this tells you about the model's fit.
10. You have a dataset with the following observed values for Y (dependent variable): 10, 20, 15, 25, and the predicted values of Y based on your model are 12, 18, 16, 22. Calculate the Mean Squared Error (MSE) of your model.
11. If the standard error of the coefficient for a predictor X in a regression model is 2.5 and the estimated coefficient is 10, calculate the t-statistic for testing the null hypothesis that the coefficient of X is zero.
12. The table below gives the monthly income and consumption expenditure of a household from a survey on consumption patterns:

t	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
C_t	29	26	35	39	46	42	60	73	54	50	69	64
Y_t	43	60	71	86	102	60	109	115	66	71	77	94

Estimate the savings function, $S_t = f(Y_t)$, and use the Spearman's Rank correlation to test for presence of heteroscedasticity in the dataset.

13. For the savings-income data below, use the Glejser's Test to check for heteroscedasticity.

Household	Savings	Income
1	2225	19271
2	2231	28528
3	1158	20720
4	1771	29848
5	409	22046
6	2222	8850
7	1100	11367
8	884	24423
9	712	14163
10	1106	16102
11	788	24098
12	936	22386
13	1623	33212
14	1384	37812
15	2073	11516
16	1387	19696
17	568	14791
18	2022	21498

Household	Savings	Income
19	1560	15557
20	439	17810
21	855	13536
22	2148	8980
23	406	26748
24	657	22995
25	1111	14175
26	1906	33338
27	1150	29899
28	1874	34361
29	1623	20686
30	826	24075

14. Following is the monthly rental of flats in a neighborhood which may or may not have a vehicle parking along with it:

Flat No.	Monthly Rent	Number of Rooms	Vehicle Parking
1	**31588**	**4**	**0**
2	**34903**	**1**	**1**
3	**24882**	**3**	**1**
4	**17355**	**5**	**1**
5	**22583**	**2**	**0**
6	**29766**	**4**	**1**
7	**14247**	**3**	**1**
8	**24354**	**2**	**1**
9	**22034**	**3**	**0**
10	**16409**	**1**	**1**

Flat No.	Monthly Rent	Number of Rooms	Vehicle Parking
11	16335	5	1
12	10385	1	0
13	13736	2	0
14	18900	2	0
15	35542	5	1
16	21660	3	1
17	22523	1	0
18	14748	2	0
19	27848	2	1
20	29357	4	1
21	30067	1	1
22	11975	1	0
23	18789	4	0
24	22914	2	0
25	32887	4	1
26	30084	1	0
27	16833	3	0
28	32738	2	0
29	16852	3	0
30	21265	1	1
31	31528	3	0

The vehicle parking is a dummy variable that takes the value 1 if the flat includes a parking spot in the apartment, and 0 if it does not. Use the data to estimate the regression equation: $Rent = b_o + b_1(Rooms) + b_2(Parking) + e$. Use the Goldfeld-Quandt Test to check for heteroscedasticity in the dataset.

15. Twenty farming lots were planted with cabbage seeds and were watered and catalyzed with urea. The yield (in tons), urea consumption (in kilograms), and water usage (in liters) is given in the table below:

Farm Lot	Yield	Urea	Water
1	58.83	17	222
2	81.58	48	70
3	61.16	7	117
4	51.18	28	294
5	67.02	18	168
6	65.4	18	266
7	62.99	43	234
8	81.79	25	67
9	97.78	11	49
10	44.12	31	214
11	86.63	35	79
12	45.73	33	287
13	59.66	34	67
14	53.38	41	113
15	42.17	4	42
16	98.06	28	87
17	45.73	18	77
18	83.91	42	67
19	99.41	17	237
20	90.54	39	234

Estimate the production function: $yield = b_0 + b_1 (urea) = b_2 (water) = e$. Test for the presence of heteroscedasticity using (a) the Spearman's Rank Correlation Test and (b) the Glejser's Test. Compare the results of the two tests.

16. For the patients in an obesity reduction program, a minimum of 1 minute of cardio exercise and walking at least 1000 steps in day was mandated. At the end of the day, a measurement of total calories burnt due to physical activity was recorded. Below is the data for 15 patients enrolled in the program:

Patient No.	Calories Burnt	Exercise Minutes	Steps Walked
1	1997	5	7747
2	1915	31	7755
3	2446	6	3186
4	2648	1	11388
5	1512	22	1527
6	2008	33	4630
7	2712	20	10786
8	2213	18	12696
9	2439	27	3478
10	2181	39	1126
11	2767	38	10368
12	2577	9	4759
13	1934	5	2578
14	2073	16	9123
15	2302	25	4751

Estimate the energy consumption function: $calorie = b_0 + b_1(exercise) + b_2(steps) + e$. Use the data to estimate the form of heteroscedasticity present.

4.1.6 References

1. **Principles of Econometrics** by R. Carter Hill, William E. Griffiths, and Guay C. Lim. This textbook provides a solid foundation in the basic principles of econometrics, including a clear explanation of

the assumptions underlying the CLRM and the importance of these assumptions for statistical inference.

2. **Econometric Methods with Applications in Business and Economics** by Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek, and Herman K. van Dijk. This book offers a comprehensive overview of econometric methods, including linear regression models, with an emphasis on applications in business and economics. It's known for its practical approach and use of case studies.
3. **Mostly Harmless Econometrics: An Empiricist's Companion** by Joshua D. Angrist and Jörn-Steffen Pischke. This book provides an intuitive approach to understanding econometric methods, including linear regression and its assumptions, focusing on how these tools are applied in empirical research. It's well-regarded for its clarity and humor, making complex concepts more accessible.
4. **Econometric Theory and Methods** by Russell Davidson and James G. MacKinnon. This textbook provides a rigorous treatment of econometric theory, including the assumptions and estimation of linear regression models. It's suitable for advanced undergraduates and graduate students seeking a deeper theoretical understanding.
5. **Econometrics by Example** by Damodar N. Gujarati. Gujarati's book is known for its clear and practical approach to econometric concepts, including the CLRM. Each chapter introduces econometric methods through real-world examples, making it an excellent resource for learners who prefer an applied perspective.

Lesson 4.2 – Nonstochastic Assumptions of the Classical Linear Regression Model

Structure

4.2.1 Introduction

The classical linear regression model relies on a set of assumptions to ensure valid statistical inference. Among these, the non-stochastic assumptions deal specifically with the independent variables (also known as regressors) in the model, denoted by the matrix $\boldsymbol{X}$. These assumptions are called "non-stochastic" because they relate to the independent variables, or the $\boldsymbol{X}$ variables, which are assumed to be fixed and not subject to random variation. The main non-stochastic assumptions in the classical linear regression model are:

1. **Correct Model Specification:** The model correctly specifies the functional form and includes all relevant variables while excluding irrelevant ones. Incorrect specification, such as omitting a relevant variable or including an irrelevant one, can lead to biased and inconsistent parameter estimates. This assumption also encompasses the correct form of the relationship (e.g., linear or log-linear) and the absence of measurement errors in the variables.

If the model is mis-specified by omitting important variables or including irrelevant ones, it can lead to biased and inconsistent estimates of the regression coefficients.

2. **Linear in Parameters:** The model is assumed to be linear in parameters, meaning the relationship between the independent variables (explanatory variables) and the dependent variable is linear. This does not mean the variables themselves must be linear; transformations can be applied. The assumption is about the parameters (coefficients), which should appear in a linear fashion in the equation.
3. **Fixed and Deterministic X:** This assumption states that the values of the independent variables in the sample are fixed and known without error. In essence, $\boldsymbol{X}$ is treated as a constant matrix. This implies there's no measurement error associated with the independent variables. It also eliminates the possibility of using lagged values of the dependent variable (y) as independent variables, which would induce serial correlation.
4. **No Endogeneity:** This assumption refers to the absence of a two-way causal relationship between the independent variables and the dependent variable. The independent variables are assumed to solely influence the dependent variable, not the other way around. Endogeneity, if present, violates this assumption. It can arise when the independent variables are correlated with the error term, leading to biased estimates. In other words, the independent variables are assumed to be determined outside the model and are not influenced by the error term. Mathematically, this can be expressed as: $\boldsymbol{E(\varepsilon \mid X_1, X_2, \ldots X_k) = 0}$. If this assumption is violated, it can lead to biased and inconsistent estimates of the regression coefficients, as the independent variables may be correlated with unobserved factors that are captured in the error term.
5. **No Multicollinearity:** This assumption ensures that the independent variables in $\boldsymbol{X}$ are not perfectly linearly dependent on each other. Perfect collinearity (where one variable can be expressed as an exact linear combination of others) would prevent the model from uniquely estimating the coefficients (β's). Even high collinearity (strong correlation between independent variables) can cause problems like inflated variances and imprecise coefficient estimates.

6. **Full-Rank X:** This assumption requires the number of observations (N) in your sample to be greater than the number of independent variables (k) in the model ($T>k$). Additionally, the rank of the $\boldsymbol{X}$ matrix must be equal to the number of independent variables (*rank* ($\boldsymbol{X}$) = k). This ensures that there are no redundant rows or exact linear relationships within the independent variables, which would again hinder the estimation process. In other words, no explanatory variable is a perfect linear combination of the others, that is, no perfect multicollinearity exists among the explanatory variables.

These non-stochastic assumptions are foundational for the classical linear regression analysis. Econometricians have developed various diagnostic tests and remedies for these issues, such as:

- Tests for linearity to check if the linear model is appropriate.
- Tests for multicollinearity, such as the Variance Inflation Factor (VIF).
- Specification tests to detect omitted variables or incorrect functional forms, like the Ramsey RESET test.
- Instrumental variables (IV) estimation or other methods to address endogeneity.

Violations of these assumptions can lead to biased and inefficient estimates, potentially rendering the inferences unreliable. We limit our discussion to only the assumptions of serial independence and multicollinearity, detecting their violations, consequences of their violations, and available remedies.

4.2.1 The Assumption of Serial Independence

Serial independence (often called lack of serial correlation or lack of autocorrelation) refers to the assumption that the error terms in a linear regression model are uncorrelated with each other over time or across observations. Mathematically, it implies that for all error terms $\grave{o}_i$ and $\grave{o}_j$ in the model, where $i \neq j$, the covariance between $\grave{o}_i$ and $\grave{o}_j$ is zero, i.e., $cov\left(\grave{o}_i, \grave{o}_j\right) = 0 \ \forall \ i \neq j$. This assumption ensures that the errors are independent across observations and that there is no pattern (like a trend or cyclical movement) in the residuals that could be predictive of other residuals. In simpler terms:

- There should be no systematic pattern in the way the error terms are related to one another.
- The error at one point in time should not provide information about the error at another point in time.

The serial independence assumption is often referred to as the "no autocorrelation" assumption because it implies that there is no correlation between the error terms across different time periods or observations.

While the key non-stochastic assumptions place strict conditions on the independent variables (***X*** matrix), the assumption of serial independence specifically deals with the behavior of the error terms (ε) in the model. But, recall that a core non-stochastic assumption is that the independent variables are fixed and predetermined. This means any randomness in the model is intended to be fully captured by the error term. If the error terms exhibit serial correlation, it means that the value of the error at one observation is influenced by errors at other points. This suggests that there's a systematic pattern or information left unexplained by the independent variables in the model. Essentially, it signals that there might be additional variables or dynamics not included in your ***X*** matrix that are affecting the dependent variable. Thus, the assumption of fixed and deterministic ***X*** matrix boils down to the assumption of no autocorrelation among the error terms.

4.2.2 Consequences of Autocorrelation

Violation of this assumption, known as serial correlation or autocorrelation, can occur in situations where the error terms are correlated over time or across observations. In time-series data or panel data, where observations are ordered sequentially, the error terms may exhibit a pattern or dependence on previous error terms. For example, economic data collected over time, such as GDP, inflation rates, or stock prices, often exhibit autocorrelation because past values influence current values.

Serial correlation can arise due to various reasons, such as omitted variables, misspecification of the model, or the presence of dynamics or persistence in the dependent variable or error terms.

The presence of serial correlation in the error terms can lead to several issues:

1. **Inefficient estimates:** The ordinary least squares (OLS) estimators, while still unbiased, are no longer efficient (minimum variance) in the presence of serial correlation.
2. **Invalid statistical inferences:** The standard errors of the estimated coefficients and hypothesis tests become unreliable, leading to invalid statistical inferences.
3. **Biased estimates:** In some cases, such as when the lagged dependent variable is included as an independent variable, serial correlation can lead to biased estimates of the coefficients.

Violations of this assumption can lead to inefficient parameter estimates and incorrect inferences about the relationship between the dependent variable and the independent variables. Therefore, it is important to test for serial independence and to apply the appropriate remedies if the assumption is violated.

4.2.3 Tests for Autocorrelation

Understanding and addressing serial independence is vital in econometric modeling to ensure the accuracy and reliability of the analysis, especially in time series data where autocorrelation is more prevalent. Plotting residuals (errors) against time or observation order can reveal patterns such as trends, cycles, or clustering suggesting serial correlation.

The simplest possible case of autocorrelation is the presence of linear correlation between two successive values in time of the error terms. Such a relationship can be represented as:

$$\dot{o}_t = \rho \dot{o}_{t-1} + \varepsilon_t$$

called the first order autoregressive relationship; ρ being the true population autocorrelation coefficient. The error term, ε, satisfies all the assumptions of the classical linear regression model:

$$E(\varepsilon) = 0$$

$$E(\varepsilon^2) = \sigma_\varepsilon^2$$

$$E(\varepsilon_i \varepsilon_j) = 0 \; \forall i \neq j$$

Therefore, if $\rho = 0$, $\dot{o}_t = \varepsilon_t$, and since ε_t is not autocorrelated, by extension, $\dot{o}_t$ is also not autocorrelated. The population autocorrelation coefficient can be estimated as:

$$r_{e_t, e_{t-1}} = \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2 \Sigma e_{t-1}^2}$$

$$= \hat{\rho}_{\text{q},\dot{\text{q}}_{-1}}$$

To detect serial correlation, econometricians typically use diagnostic tests, such as:

The Von Neumann Ratio

The von Neumann Ratio test is a statistical test used to detect the presence of first-order autocorrelation (serial correlation) in time series data or within the residuals of a regression model. The Von Neumann Ratio test is based on the ratio of the mean square successive difference to the variance of the time series. The test statistic is given by:

$$\frac{\delta^2}{S_x^2} = \frac{\frac{\Sigma_{t=2}^{n} \left(X_t - X_{t-1}\right)^2}{n-1}}{\frac{\Sigma \left(X_t - \bar{X}\right)^2}{n}}$$

The ratio is calculated applicable to directly observed non-autocorrelated data in a series. Since the error terms are not directly observable, their estimates can be used to approximate the ratio as:

$$\frac{\delta^2}{S_e^2} = \frac{\frac{\Sigma_{i=2}^{n} \left(e_t - e_{t-1}\right)^2}{n-1}}{\frac{\Sigma_{i=1}^{n} \left(e_t - \bar{e}\right)^2}{n}}$$

Where:

- e_t : residual at time t
- $e_{(t-1)}$: Residual at time $t-1$
- $\bar{e}$: Mean of the residuals

The Von Neumann ratio tests for the randomness of a sequence, where randomness implies the absence of autocorrelation. The logic is that if there's no autocorrelation, successive differences between residuals should be random. If there's positive autocorrelation, successive differences will tend to be small (as positively correlated values stay close to each other, successive observations are similar). If there's negative autocorrelation, successive differences will tend to be large (that is, successive observations differ significantly) with values oscillating.

The test statistic of the Von Neumann ratio test, under the null hypothesis of independence (i.e., no autocorrelation), asymptotically follows a normal distribution for large sample sizes. Specifically, when the number of observations n is large, the distribution of the test statistic approaches normality due to the Central Limit Theorem. This property allows for the derivation of critical values and the assessment of statistical significance based on the normal distribution.

However, for the Von Neumann ratio specifically, the expected value under the null hypothesis of no autocorrelation is close to 2, and its variance can be expressed as a function of the sample size n. For practical purposes, when using this test, one often refers to tabulated values or specific statistical software to determine the critical values for the test statistic that correspond to conventional levels of statistical significance (e.g., $\alpha = 0.05$).

While the normal distribution approximation works well for large sample sizes, for smaller samples, the exact distribution of the test statistic might need to be considered, and the critical values can be different from those suggested by the normal approximation. The use of simulation or bootstrap methods can also help in assessing the distribution of the test statistic and its critical values for smaller samples or when a high level of accuracy is required. The von Neumann Ratio will generally be between 0 and 4.

- A value close to 2 indicates no autocorrelation.
- A value below 2 suggests positive autocorrelation.
- A value above 2 suggests negative autocorrelation.

The von Neumann Ratio test is primarily designed to detect first-order autocorrelation. It might not be as powerful in detecting higher-order autocorrelation or more complex patterns. The Durbin-Watson test is a commonly used alternative that is closely related to the von Neumann Ratio test.

The Durbin-Watson Test

The Durbin-Watson test, named after James Durbin and Geoffrey Watson, is a statistical test used to detect the presence of **first-order autocorrelation** (serial correlation) in the residuals of a linear regression model. Both Von Neumann Ratio test and the Durbin-Watson test

address the same issue: detecting first-order autocorrelation. They are mathematically related, with the Durbin-Watson statistic being a transformation of the von Neumann Ratio. The Durbin-Watson test is more widely used due to its simpler interpretation and readily available critical values. The Durbin-Watson statistic (DW) is calculated as follows:

$$DW = \frac{\Sigma_{t=2}^{n}\left(e_t - e_{(t-1)}\right)^2}{\Sigma_{t=1}^{n}\left(e_t - \overline{e}\right)^2}$$

Where:

- e_t: residual at time t
- $e_{(t-1)}$: Residual at time $t-1$
- $\overline{e}$: Mean of the residuals

Thus, the Von Neumann Ratio and the Durbin-Watson test statistic are related as:

$$DW = \left(\frac{\delta^2}{S^2}\right)\left(\frac{n-1}{n}\right)$$

$$= \left(\frac{\frac{\Sigma_{i=2}^{n}\left(e_t - e_{t-1}\right)^2}{n-1}}{\frac{\Sigma_{i=1}^{n}\left(e_t - \overline{e}\right)^2}{n}}\right)\left(\frac{n-1}{n}\right)$$

$$= \frac{\Sigma_{t=2}^{n}\left(e_t - e_{(t-1)}\right)^2}{\Sigma_{t=1}^{n}\left(e_t - \overline{e}\right)^2}$$

For large sample size, $n \approx n-1$, thus the Durbin-Watson test statistic tends to the Von Neumann ratio asymptotically. The DW statistic ranges from 0 to 4; a value of ***2*** implies no autocorrelation ($\hat{\rho} = 0$), while $DW < 2$ suggests positive autocorrelation (residuals tend to be positively correlated with their neighbors), and $DW > 2$ suggests negative autocorrelation (residuals tend to be negatively correlated with their neighbors). That is:

$$2 \le DW \le 4 \equiv -1 \le \hat{\rho} \le 1$$

The Durbin-Watson statistic does not follow a simple statistical distribution. Instead, we rely on critical values provided by Durbin and Watson in statistical tables or by a statistical software. These critical values depend on the sample size and the significance level chosen (usually 0.05 or 0.01). There are two sets of critical values:

- **Lower Critical Value (DW_L)**
- **Upper Critical Value (DW_U)**

We compare DW with the critical values and if:

- $DW > DW_U$: we do not reject the null hypothesis (no autocorrelation).
- $DW < DW_L$: we do not accept the null hypothesis (evidence of positive autocorrelation).
- $DW_L < DW < DW_U$: then the test is inconclusive (weak evidence for autocorrelation, may need a larger sample size or alternative tests).

The Durbin-Watson test is applicable to linear regression models where the residuals are assumed to be independently and identically distributed (i.i.d.). It may not be as sensitive to higher-order autocorrelation or more complex patterns. The test is particularly useful in time series analysis where autocorrelation is a common concern. The interpretation of the DW statistic relies on critical values, which can be affected by factors like sample size and the presence of constant terms in the regression model. The test is not suitable for models with lagged dependent variables as an explanatory variable. In such cases, alternative tests like Durbin's *h*-test are more appropriate. In case of higher order correlations, a more general test, such as the Breusch-Godfrey Test, is more appropriate.

Breusch-Godfrey Test

The Breusch-Godfrey Test is an extension of the Durbin-Watson test, which only detects first-order autocorrelation. The Breusch-Godfrey Test can detect autocorrelation up to any specified order. To calculate the test statistic, we estimate the original linear regression model and obtain the residuals and then specify the order of autocorrelation to be tested (p). We, then regress the residuals on the original regressors and their lagged values up to order p and calculate the test statistic using the formula:

$$BG = (n - p - 1) R^2 \sim \chi^2 (p)$$

where:

- n is the sample size,
- p is the order of autocorrelation,

- R^2 is the coefficient of determination from the auxiliary regression, and
- $\chi^2(p)$ is the chi-squared distribution with p degrees of freedom.

The decision criteria for the Breusch-Godfrey Test is as follows: if the test statistic BG is greater than the critical value from the chi-squared distribution with p degrees of freedom, reject the null hypothesis (H_0) of no autocorrelation; if the test statistic BG is less than or equal to the critical value, fail to reject the null hypothesis (H_0) of no autocorrelation.

Alternatively, an F-test can be applied to assess the joint significance of the lagged residuals in the auxiliary regression. The formula for the F-test statistic is:

$$F = \frac{\frac{(RSS_0 - RSS_1)}{p}}{\frac{RSS_1}{(n-k-p)}}$$

where:

- RSS_0 is the residual sum of squares from the original regression model without lagged residuals,
- RSS_1 is the residual sum of squares from the auxiliary regression model that includes the lagged residuals,
- p is the number of lagged residual terms (degrees of freedom for the numerator),
- n is the sample size,
- k is the number of independent variables in the original model (excluding the constant term and the lagged residuals), and
- $(n–k–p)$ is the degrees of freedom for the denominator.

This F-test statistic is used to determine if the lagged residuals are jointly significant, indicating the presence of serial correlation. This is suitable for small sample sizes. If the calculated F-statistic is greater than the critical value from the F-distribution for the given degrees of freedom and significance level, the null hypothesis of no serial correlation is rejected.

The Breusch-Godfrey Test is widely applicable in time-series analysis and other situations where autocorrelation is a concern. However, it has some limitations:

- It assumes that the error terms are normally distributed, which may not be true in all cases.
- It is sensitive to the choice of the order of autocorrelation, *p*. If *p* is chosen incorrectly, the test may not detect existing autocorrelation.
- The test may have low power in small samples, making it difficult to detect autocorrelation when it exists.

Despite these limitations, the Breusch-Godfrey Test remains a popular and useful tool for detecting autocorrelation in linear regression models.

Ljung-Box Test

The Ljung-Box test (also known as the Ljung-Box Q test or the modified Box-Pierce test) is a statistical method used to investigate the presence of autocorrelation (serial correlation) in a time series or in the residuals of a fitted model.

The Ljung-Box test helps determine whether the residuals of a model exhibit autocorrelation. The basic idea is that if there is no significant autocorrelation, the residuals should be random. If there is significant autocorrelation, there is likely a pattern in the residuals, and the model has not fully captured all the structure in the data.

The Ljung-Box tests the null hypothesis that the data is independently distributed (no serial correlation) against the alternative hypothesis of serial correlation up to a specified lag. It is also used to determine the appropriate lag order for an autoregressive moving average (ARMA) model or other time series models. The test statistics is calculated as:

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k}$$

where:

- n is the sample size,
- h is the number of lags being tested,
- $\hat{\rho}_k$ is the sample autocorrelation at lag k.

This test statistic follows a χ^2 distribution with h degrees of freedom if the null hypothesis is true. The Ljung-Box test is applicable for checking the autocorrelation in the residuals of a fitted ARIMA model, for instance. It is suitable for continuous data and is commonly used in the field of finance, meteorology, and hydrology for time series analysis.

But the Ljung-Box test may have low power when applied to data with non-normal distributions. The test is not suitable for very small sample sizes because it relies on asymptotic distribution, and it can be overly sensitive to departures from the null hypothesis when the sample size is large. The test may also not be appropriate for models that include lagged dependent variables or models with parameter estimates that are close to the non-stationary boundary.

In practice, it is important to interpret the results of the Ljung-Box test in conjunction with other tests and model diagnostics, and not to rely solely on this test when assessing the adequacy of a time series model.

4.2.4 Estimation with Autocorrelation

When dealing with autocorrelated data, there are several solutions that can be employed to address this issue. The choice of solution depends on the nature of the data, the underlying model, and the specific research or modeling objectives. Here are some common solutions for autocorrelated data:

Autoregressive Models (AR): If the autocorrelation pattern is systematic and can be modeled, autoregressive models (AR) can be used. In an AR model, the current value of the dependent variable is modeled as a function of its past values and an error term. This approach is applicable when the autocorrelation is a result of the time series structure itself. The models assumes that the time series is stationary, and the autocorrelation pattern can be adequately captured by the AR model. If the autocorrelation structure is mis-specified or if there are structural breaks or nonlinearities, the AR model may not be appropriate.

Moving Average Models (MA): Moving average models (MA) can be used when the autocorrelation is caused by random shocks or disturbances. In an MA model, the current value of the dependent variable is modeled as a function of past error terms. This approach is suitable when the autocorrelation is due to external factors or interventions. The effectiveness of the model lies on the assumption of the autocorrelation being caused by random shocks or disturbances, and the error process following a specific MA structure. The MA models may not be suitable if the autocorrelation is systematic or if the error process is more complex than the assumed MA structure.

Autoregressive Integrated Moving Average (ARIMA): ARIMA models combine autoregressive and moving average components, making them suitable for handling both systematic and random autocorrelation patterns. These models are applicable when the time series is stationary or can be made stationary through differencing. The model assumes that the time series is stationary or can be made stationary through differencing, and the autocorrelation pattern can be captured by the ARIMA model. These models may not be appropriate if the time series exhibits nonlinearities, structural breaks, or if the autocorrelation structure is more complex than the assumed ARIMA model.

Generalized Autoregressive Conditional Heteroskedasticity (GARCH): GARCH models are used when the autocorrelation is present in the variance (heteroskedasticity) rather than the mean of the time series. These models are useful for modeling volatility in financial time series data and assume that the autocorrelation is present in the variance (heteroskedasticity) of the time series, and that the conditional variance follows a specific GARCH structure. GARCH models may not be suitable if the heteroskedasticity pattern is more complex than the assumed GARCH structure or if there are other forms of non-stationarity in the data.

Generalized Least Squares (GLS): If the autocorrelation pattern is known or can be estimated, the GLS method can be used to obtain efficient and unbiased estimates of the regression coefficients. This approach involves transforming the original model to account for the autocorrelation structure. The GLS approach is based on the assumption that the autocorrelation pattern is known or can be estimated accurately, and the error process follows a specific structure (e.g., AR or MA). But, if the autocorrelation structure is mis-specified or if the assumptions of the GLS method are violated, the resulting estimates may be biased or inefficient.

Cochrane-Orcutt Procedure: The Cochrane-Orcutt procedure is an iterative method for estimating regression coefficients in the presence of first-order autocorrelation. It involves transforming the original model and then applying ordinary least squares (OLS) to the transformed model. The model assumes that the autocorrelation follows a first-order autoregressive process, and the error process is

serially uncorrelated after the transformation. In this method the autocorrelation coefficient, ρ, is iteratively estimated at each stage of transformation till the estimated values converge. For example, as first step, the OLS procedure is run to compute the residuals and estimate the coefficient of autocorrelation as:

$$\hat{\rho} = \frac{\Sigma e_t e_{t-1}}{\Sigma e_{t-1}^2}$$

This value is then used to transform the original regression model and its underlying data as:

$$\left[Y_t - \hat{\rho} Y_{t-1}\right] = \beta_0 \left(1 - \hat{\rho}\right) + \beta_1 \left[X_t - \hat{\rho} X_{t-1}\right] + \text{ὸ}_t^*$$

and then the OLS procedure is applied to this transformed data to arrive at second stage error estimates. These values are then used to estimates the second stage coefficient of autocorrelation as:

$$\hat{\hat{\rho}} = \frac{\Sigma \hat{\hat{e}}_t \hat{\hat{e}}_{t-1}}{\Sigma \hat{\hat{e}}_{t-1}^2}$$

The $\hat{\hat{\rho}}$ is used to transform the original data as:

$$\left[Y_t - \hat{\hat{\rho}} Y_{t-1}\right] = \beta_0 \left(1 - \hat{\hat{\rho}}\right) + \beta_1 \left[X_t - \hat{\hat{\rho}} X_{t-1}\right] + \text{ὸ}_t^{**}$$

OLS is run again to estimate third stage coefficient of autocorrelation as:

$$\hat{\hat{\hat{\rho}}} = \frac{\Sigma \hat{\hat{\hat{e}}}_t \hat{\hat{\hat{e}}}_{t-1}}{\Sigma \hat{\hat{\hat{e}}}_{t-1}^2}$$

This procedure (iterations) is repeated till the estimates of the coefficient of autocorrelation at different stages $(\rho, \hat{\rho}, \hat{\hat{\rho}}, \hat{\hat{\hat{\rho}}}, \ldots)$ converge. The most efficient prediction of the outcome variable in the presence of autocorrelation can be done by:

$$\hat{Y}_{n+1} = \beta_0^* + \beta_1^* X_{1(n+1)} + \beta_2^* X_{2(n+1)} + \ldots + \beta_k^* X_{k(n+1)} + \rho^* e_n$$

where:

e_n = the error of the n^{th} observation in the final model,

ρ^* = the estimate of the coefficient of autocorrelation obtained from the final iteration,

$\beta^* s$ = the estimates of the regression parameters using the transformed variables of the final model in the final iteration.

The Cochrane-Orcutt procedure may not be appropriate if the autocorrelation structure is more complex than a first-order autoregressive process or if there are other violations of the assumptions (e.g., heteroskedasticity).

Newey-West Standard Errors: When the autocorrelation structure is unknown or difficult to model, the Newey-West standard errors can be used to obtain robust standard errors for the regression coefficients, accounting for potential autocorrelation and heteroskedasticity. It assumes that while the error process may exhibit autocorrelation and heteroskedasticity, but the underlying model is correctly specified. But, if the model is mis-specified or if there are other violations of the assumptions (e.g., non-stationarity), the Newey-West standard errors may not provide reliable inferences.

To make use of these solutions, it is important to first diagnose the presence and nature of autocorrelation using appropriate statistical tests (e.g., Durbin-Watson, Breusch-Godfrey, or Ljung-Box tests). Once the autocorrelation pattern is identified, the appropriate solution can be selected based on the characteristics of the data and the underlying model assumptions.

For example, if the autocorrelation is due to an omitted variable, the appropriate solution would be to incorporate the omitted variable in the model. The instance of a 'quasi'-autocorrelation can be illustrated by an example: a person's savings in the current period (t) does not only depend on their current income (Y_t), but also on their income levels from previous periods. This can be represented in a simple savings function as:

$$S(t) = f(Y_t, Y_{t-1})$$

$$S_t = \beta_0 + \beta_1 Y_t + \beta_2 X_{t-1} + \text{ò}_t$$

Where Y_{t-1} is the income from the previous period. If this lagged income term (Y_{t-1}) is omitted from the savings function, its effect will get absorbed into the random error term (ε_t). Additionally, the coefficient estimate for current income (Y_t) is likely to be biased due to this omission.

Since income values are usually positively correlated over time, omitting the lagged income term will result in a pattern of autocorrelation in the error terms (ε's) across different time periods. This autocorrelation

issue can be resolved by explicitly including the lagged income term as an explanatory variable in the savings function.

To detect if autocorrelation is caused by omitting relevant variables, one approach is to regress the residuals (*e*'s) from the initial model on variables that could potentially explain the phenomenon being studied based on theory or prior expectations. If such variables are significant in explaining the residuals, it suggests they should be included in the original model to address omitted variable bias and autocorrelation.

Thus, we must carefully examine the assumptions and limitations of each solution before applying it to a specific dataset or modeling problem. In some cases, a combination of different approaches or more advanced techniques (e.g., state-space models, non-parametric methods) may be required to effectively address autocorrelation, particularly when the underlying data generating process is complex or exhibits non-linear behavior.

4.2.5 The Assumption of Non-multicollinear Regressors

The assumption of non-multicollinear regressors, also known as the absence of perfect multicollinearity, is one of the fundamental assumptions in multiple linear regression analysis. This assumption is crucial for the proper estimation and interpretation of the regression coefficients.

Multicollinearity refers to the existence of a linear relationship or correlation among the independent variables (regressors) in a multiple regression model. When multicollinearity is present, it becomes difficult to disentangle the individual effects of each independent variable on the dependent variable, as the independent variables are highly correlated with one another.

The assumption of non-multicollinear regressors implies that the independent variables in the regression model should not be perfectly correlated with one another. In other words, there should not be an exact linear relationship among the independent variables. It is important to note that the assumption of non-multicollinear regressors does not require the complete absence of correlation among the independent variables. Some degree of correlation is acceptable and often expected in real-world data. However, the assumption is violated when there is a perfect or near-perfect linear relationship among the independent variables.

To detect multicollinearity, researchers often examine the correlation matrix of the independent variables, calculate variance inflation factors (VIFs), or employ other diagnostic measures. If multicollinearity is detected, several remedial measures can be taken, such as removing one or more highly correlated variables from the model, combining correlated variables into a single variable, or employing techniques like ridge regression or principal component regression.

By satisfying the assumption of non-multicollinear regressors, the regression model can provide reliable and interpretable estimates of the individual effects of each independent variable on the dependent variable, allowing for meaningful statistical inference and analysis.

4.2.5 Consequences of Multicollinearity

The presence of multicollinearity in a multiple regression model can have several consequences. Perfect correlation among the predictors ($r_{x_i x_j} = 1$) can affect the reliability and interpretability of the regression model being investigated:

1. **Infinitely large standard errors of estimates:** Multicollinearity can lead to large standard errors for the regression coefficients, making it difficult to assess their statistical significance accurately. This is because multicollinearity leads to an increase in the variance of the estimated regression coefficients, making them less precise and less reliable. The variance of the ordinary least squares (OLS) estimator for the regression coefficient, $\hat{\beta}_j$, is given by:

$$\text{var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\left[\left(1 - R_j^2\right)\Sigma\left(x_{ij} - x_j\right)^2\right]}$$

where:

σ^2 is the variance of the error term,

R_j^2 is the coefficient of determination when x_j is regressed on the remaining independent variables,

$\Sigma\left(x_{ij} - x_j\right)^2$ is the sum of squared deviations of x_j from its mean,

As R_j^2 approaches 1 (indicating a high degree of multicollinearity), the denominator decreases, leading to an increase in the variance of $\hat{\beta}_j$. Since the standard error of $\hat{\beta}_j$ is given by:

$$SE\left(\hat{\beta}_j\right) = \sqrt{var\left(\hat{\beta}_j\right)}$$

multicollinearity increases the variance of $\hat{\beta}_j$, leading to larger standard errors and wider confidence intervals for the regression coefficients. In the event of perfect multicollinearity, calculating the variance entails division by zero resulting in:

$$var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{0} = \infty$$

$$SE\left(\hat{\beta}_j\right) = \sqrt{\infty} = \infty$$

2. **Indeterminate estimation of regression coefficients:** Perfect multicollinearity renders the estimates of the regression coefficients indeterminate. Consider the regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{ò}$$

Where the independent variables are perfectly correlated as: $X_2 = kX_1$; k being an arbitrary constant. Now, the equations that give the estimates of the regression parameters are:

$$b_1 = \frac{\left(\Sigma x_1 y\right)\left(\Sigma x_2^2\right) - \left(\Sigma x_2 y\right)\left(\Sigma x_1 x_2\right)}{\left(\Sigma x_1^2\right)\left(\Sigma x_2^2\right) - \left(\Sigma x_1 x_2\right)^2}$$

$$b_2 = \frac{\left(\Sigma x_2 y\right)\left(\Sigma x_1^2\right) - \left(\Sigma x_1 y\right)\left(\Sigma x_1 x_2\right)}{\left(\Sigma x_1^2\right)\left(\Sigma x_2^2\right) - \left(\Sigma x_1 x_2\right)^2}$$

Substituting $X_2 = kX_1$, we get:

$$b_1 = \frac{k^2\left(\Sigma x_1 y\right)\left(\Sigma x_1^2\right) - k^2\left(\Sigma x_1 y\right)\left(\Sigma x_1^2\right)}{k^2\left(\Sigma x_1^2\right)^2 - k^2\left(\Sigma x_1^2\right)^2} = \frac{0}{0}$$

$$b_2 = \frac{k\left(\Sigma x_1 y\right)\left(\Sigma x_1^2\right) - k\left(\Sigma x_1 y\right)\left(\Sigma x_1^2\right)}{k^2\left(\Sigma x_1^2\right)^2 - k^2\left(\Sigma x_1^2\right)^2} = \frac{0}{0}$$

Thus, the parameters of the regression model cannot be estimated and their individual contributions to the dependent variable (Y) become difficult to isolate and interpret accurately.

3. **Misleading interpretation of regression coefficients:** In the presence of multicollinearity, the regression coefficients can have counterintuitive signs or magnitudes, leading to misleading interpretations. This can happen when the independent variables are highly correlated and the researcher decides to avoid multicollinearity by omitting one of the variables. For example, let us assume that the true regression model is given by:

$$y = b_1 x_1 + b_2 x_2 + \text{ò}$$

But the researcher removes x_2 from the equation dues to its high correlation with x_1, then the relationship he estimates becomes:

$$y = b_1^* x_1 + \varepsilon$$

He applies the OLS to this functional relationship to obtain the parameter estimate as:

$$b_1^* = \frac{\Sigma y x_1}{\Sigma x_1^2}$$

Now, for the true regression relationship, the normal equations are:

$$\Sigma y x_1 = b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$
$$\Sigma y x_2 = b_2 \Sigma x_2^2 + b_1 \Sigma x_1 x_2$$

We divide the first equation with Σx_1^2 to obtain:

$$\frac{\Sigma y x_1}{\Sigma x_1^2} = b_1 + b_2 \frac{\Sigma x_1 x_2}{\Sigma x_1^2}$$

But from the earlier result, we have: $\frac{\Sigma y x_1}{\Sigma x_1^2} = b_1^*$, therefore:

$$b_1^* = b_1 + b_2 \frac{\Sigma x_1 x_2}{\Sigma x_1^2}$$

Thus, by omitting variable x_2, the researcher overestimates b_1 and we define the specification bias as:

$$\left[\text{specification bias}\right] = \left[E\left(b_1^*\right) - b_1 \right]$$
$$= b_2 \frac{\Sigma x_1 x_2}{\Sigma x_1^2}$$

Since the true regression model specifies a b_2, the specification bias will be zero (eliminated) only when:

$$\frac{\Sigma x_1 x_2}{\Sigma x_1^2} = 0$$

And this can happen only when x_1 and x_2 are perfectly non-correlated (orthogonal). Thus, multicollinearity will in general lead to a specification bias in the parameter estimates.

It is important to note that while multicollinearity does not render the OLS estimator biased, that is $E\left(b_i\right) = \beta_i$ holds, it can lead to unreliable and unstable estimates of the regression coefficients,

making it difficult to draw meaningful inferences from the regression analysis.

4.2.7 Tests for Multicollinearity

There are several statistical tests and diagnostics used to detect multicollinearity in a multiple regression model:

Frisch's Confluence Analysis

Also known as the Frisch Method or the Frisch Source Components Technique, is a method used to identify the sources of multicollinearity in a multiple regression model. It was developed by the economist Ragnar Frisch in the 1930s.

The basic idea behind Frisch's Confluence Analysis is to decompose the regressors (independent variables) into two components: one component that is free from multicollinearity and another component that captures the multicollinearity among the regressors. This decomposition allows for the identification of the specific regressors or combinations of regressors that are responsible for the multicollinearity problem. The steps involved in Frisch's Confluence Analysis are as follows:

1. Center the data: Subtract the mean from each regressor to obtain mean-centered regressors.
2. Calculate the correlation matrix ($\boldsymbol{R}$) for the mean-centered regressors.
3. Find the eigenvalues (λ_i) and corresponding eigenvectors (e_i) of the correlation matrix $\boldsymbol{R}$.
4. Arrange the eigenvectors in descending order of their corresponding eigenvalues.
5. Construct the Frisch source components (F_i) by multiplying the mean-centered regressors with the eigenvectors: $F_i = \boldsymbol{X}e_i$, where $\boldsymbol{X}$ is the matrix of mean-centered regressors, and e_i is the i^{th} eigenvector.
6. Identify the source components (F_i) with small eigenvalues (λ_i) as the sources of multicollinearity.
7. Analyze the coefficients of the problematic source components to determine the regressors contributing to multicollinearity.

Let's consider a multiple regression model with three regressors: X_1, X_2, and X_1. We first center the data:

$$X_1^{centered} = X_1 - \bar{X}_1$$

$$X_2^{centered} = X_2 - \bar{X}_2$$

$$X_3^{centered} = X_3 - \bar{X}_3$$

We then calculate the correlation matrix (***R***). Suppose the matrix turns out to be:

$$\boldsymbol{R} = \begin{bmatrix} 1.0000 & 0.9500 & 0.9000 \\ 0.9500 & 1.0000 & 0.9500 \\ 0.9000 & 0.9500 & 1.0000 \end{bmatrix}$$

Next, we find the eigenvalues and eigenvectors of ***R***: $\lambda_1 = 2.8500$, $\lambda_2 = 0.1000$, and $\lambda_3 = 0.0500$. Then the eigenvectors are:

$$e_1 = \begin{bmatrix} 0.5774 & 0.5774 & 0.5774 \end{bmatrix}$$

$$\mathrm{e}_2 = \begin{bmatrix} -0.7071 & 0.0000 & 0.7071 \end{bmatrix}$$

$$\mathrm{e}_3 = \begin{bmatrix} -0.4082 & 0.8165 & -0.4082 \end{bmatrix}$$

Next, we arrange the eigenvectors in descending order of their corresponding eigenvalues and construct the Frisch source components as:

$$F_1 = X_1^{centered}\left(0.5774\right) + X_2^{centered}\left(0.5774\right) + X_3^{centered}\left(0.5774\right)$$

$$\mathrm{F}_2 = \mathrm{X}_1^{\mathrm{centered}}\left(-0.7071\right) + \mathrm{X}_2^{\mathrm{centered}}\left(0.0000\right) + \mathrm{X}_3^{\mathrm{centered}}\left(0.7071\right)$$

$$\mathrm{F}_3 = \mathrm{X}_1^{\mathrm{centered}}\left(-0.4082\right) + \mathrm{X}_2^{\mathrm{centered}}\left(0.8165\right) + \mathrm{X}_3^{\mathrm{centered}}\left(-0.4082\right)$$

The source components F_2 and F_3 have small eigenvalues ($\lambda_2 = 0.1000 \text{and } \lambda_3 = 0.0500$), indicating that they are the sources of multicollinearity. For F_2, X_1 and X_3 have non-zero coefficients (–0.7071 and 0.7071, respectively), suggesting that X_1 and X_3 are contributing to multicollinearity. For F_3, X_2 has a large coefficient (0.8165), indicating that X_2 is also contributing to multicollinearity. Thus, based on the Frisch's Confluence Analysis, we can conclude that the regressors X_1, X_2, and X_3 all contribute to the multicollinearity problem in this regression model.

Frisch's Confluence Analysis provides a systematic approach to identifying the sources of multicollinearity and can be particularly useful when dealing with a large number of regressors. However, it is important to

note that the interpretation of the results may not always be straightforward, especially when there are multiple sources of multicollinearity present.

The Farrar-Glauber Test

The Farrar-Glauber Test is a statistical test used to detect multicollinearity in a multiple regression model. It is based on the principle of testing the significance of the regression coefficients when individual regressors are removed from the model. The Farrar-Glauber Test involves the following steps:

1. Fit the full regression model with all the regressors included.
2. For each regressor, fit a reduced regression model by excluding that regressor from the full model.
3. Calculate the F-statistic for the test of the joint significance of the excluded regressor(s) in the reduced model.
4. Compare the calculated F-statistic with the critical *F*-value from the *F*-distribution table at the chosen significance level.

The null hypothesis for the Farrar-Glauber Test is that there is no multicollinearity among the regressors, while the alternative hypothesis is that multicollinearity exists. The test statistic for the Farrar-Glauber Test is calculated as follows:

$$F = \frac{RSS_{reduced} - RSS_{full}}{\frac{q \times RSS_{full}}{n-k}}$$

Where:

$RSS_{reduced}$ is the residual sum of squares from the reduced model (without the regressor(s) being tested),

RSS_{full} is the residual sum of squares from the full model (with all regressors),

q is the number of regressors excluded from the reduced model,

n is the number of observations, and

k is the number of regressors in the full model.

If the calculated *F*-statistic exceeds the critical *F*-value from the *F*-distribution table at the chosen significance level, the null hypothesis is rejected, indicating the presence of multicollinearity among the regressors.

Consider a multiple regression model with three regressors: X_1, X_2, and X_3, and a dependent variable Y. Suppose we have the following regression results (full model):

$$Y = 2.5 + 0.8X_1 + 0.6X_2 + 0.4X_3$$

$$RSS_{full} = 100 \quad n = 50 \quad k = 4\left(\text{including the intercept}\right)$$

Reduced Model 1 (excluding X_1):

$$Y = 1.8 + 0.7X_2 + 0.5X_3$$

$$\text{RSS}^1_{\text{reduced}} = 120$$

Reduced Model 2 (excluding X_2):

$$Y = 2.2 + 0.9X_1 + 0.3X_3$$

$$\text{RSS}^2_{\text{reduced}} = 110$$

Reduced Model 3 (excluding X_3):

$$Y = 2.1 + 0.9X_1 + 0.7X_2$$

$$\text{RSS}^3_{\text{reduced}} = 105$$

To perform the Farrar-Glauber Test for each regressor, we calculate the *F*-statistic and compare it with the critical *F*-value at the chosen significance level (e.g., α=5%):

$$F_{X_1} = \frac{\left(RSS^1_{reduced} - RSS_{full}\right)}{\frac{\left(1 \times RSS_{full}\right)}{\left(50-4\right)}}$$

$$= \frac{\left(120-100\right)}{\frac{1\times 100}{46}}$$

$$= 8.696$$

$$F_{X_2} = \frac{\left(RSS^2_{reduced} - RSS_{full}\right)}{\frac{\left(1 \times RSS_{full}\right)}{\left(50-4\right)}}$$

$$= \frac{\left(110-100\right)}{\frac{1\times 100}{46}}$$

$$= 4.348$$

$$F_{X_3} = \frac{\left(RSS^3_{reduced} - RSS_{full}\right)}{\frac{\left(1 \times RSS_{full}\right)}{\left(50-4\right)}}$$

$$= \frac{\left(105-100\right)}{\frac{1\times 100}{46}}$$

$$= 2.174$$

The $F_{critical}$ value from the *F*-distribution table with 1 and 46 degrees of freedom at the 5% significance level is approximately 4.05. Since the $F_{X_1} > F_{critical}$, we reject the null hypothesis and conclude that multicollinearity exists among the regressors, potentially involving X_1.

But, although $F_{X_2} > F_{critical}$, the value close enough to the critical *F*-value, suggesting that multicollinearity may exist, but the evidence is weaker compared to X_1. It is only with X_3, since $F_{X_3} < F_{critical}$, that we may not reject the null hypothesis and conclude that multicollinearity in the dataset does not involve X_3.

The Farrar-Glauber Test provides a way to identify multicollinearity by examining the significance of the regression coefficients when individual regressors are removed from the model. However, it should be noted that the test may not always be conclusive, especially when multicollinearity is present among multiple regressors simultaneously. Additionally, the Farrar-Glauber Test relies on the assumption of normality and homoscedasticity of the error terms, and violations of these assumptions may affect the validity of the test results.

Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) is a measure of how much the variance of an estimated regression coefficient increases if your predictors are correlated. If X_i is one of the predictors, the VIF for X_i is calculated as:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the coefficient of determination of a regression of X_i on all the other predictors. A common rule of thumb is that if $VIF > 10$, then multicollinearity is high. Some researchers also use a more conservative threshold of 5.

The VIF method assumes a linear relationship between the predictor of interest and other predictors and can only assess multicollinearity on a variable-by-variable basis and may not detect multicollinearity arising from a combination of variables.

Tolerance

Tolerance is the inverse of VIF and is calculated as:

$$\tau_i = 1 - R_i^2$$

Tolerance values lower than 0.1 (or some say 0.2) suggest multicollinearity may be influencing the regression estimates. Its assumptions are similar to that of the VIF—assumes linear relationships and considers variables individually.

Condition Index

The Condition Index is used to assess the severity of multicollinearity. It's based on the singular value decomposition of the scaled, centered design matrix $\boldsymbol{X}$. The condition number for each dimension is calculated as:

$$\text{Condition Index} = \frac{\text{Largest Eigenvalue}}{\text{Eigenvalue}_i}$$

A condition index greater than 30 is often taken as a sign of strong multicollinearity. But the condition index does not identify which variables are causing multicollinearity—high condition index might be caused by a single predictor or a combination of predictors.

Eigenvalues of the Correlation Matrix

If the correlation matrix of the predictors is denoted by ***R***, multicollinearity exists if one or more of the eigenvalues of ***R*** are close to zero. If any eigenvalue is close to 0, the corresponding condition index will be high, indicating multicollinearity. This method is limited by the fact that it relies on the correlation matrix, which considers linear associations, therefore, it cannot assess non-linear collinearity.

Correlation Matrix

The correlation matrix itself can be inspected to detect multicollinearity by looking at the pairwise correlation coefficients between independent variables. A high correlation coefficient (e.g., above 0.8 or below -0.8) between two or more predictors suggests multicollinearity. This method may is not helpful in detecting multicollinearity if it is caused by a combination of three or more variables.

In general, these tests are indicative rather than definitive. They can suggest the presence of multicollinearity but do not always identify its specific nature or the best course of action. The thresholds used for decision criteria are somewhat arbitrary and should be used as guidelines rather than hard rules. Also, these tests do not account for non-linear multicollinearity. Moreover, multicollinearity is not always a problem *per se*; it only becomes problematic when we need precise estimates of individual predictor effects or if we are trying to avoid overfitting in predictive models.

These tests can help diagnose the presence of multicollinearity, but the interpretation of the results should take into account the context of the study and the goals of the analysis. When multicollinearity is detected, it may be addressed through methods such as dropping variables, combining variables, or using ridge regression, which is a regularization technique that can handle multicollinearity.

4.2.8 Solutions for Multicollinear Data

Dealing with multicollinearity in a dataset involves either removing the multicollinearity or mitigating its impact on the model. Following are some common methods employed:

Increasing Sample Size: With a larger sample size, the estimates become more stable, and the impact of multicollinearity is lessened. But collecting more data can be expensive or time-consuming, and it might not always be feasible.

Removing correlated variables: If two variables are highly correlated, consider removing one of them from the model. This simplifies the model by eliminating variables that provide redundant information. This approach is simple but can be problematic as it can potentially lose important information that the removed variable(s) could have explained.

Linear combinations: Create a new variable that is a linear combination (e.g., sum or average) of the correlated variables. This can help reduce multicollinearity but may not be appropriate if the original variables have different interpretations or are measured on different scales.

Regularization techniques: Ridge Regression and Lasso are two techniques that can be used to address multicollinearity. Both methods add a penalty term to the loss function to shrink the coefficients and reduce the impact of multicollinearity.

Ridge Regression minimizes the following objective function:

$$min_{\beta} \Sigma\left(y_i - \beta_0 - \Sigma\left(x_{ij} * \beta_j\right)\right)^2 + \lambda * \Sigma\left(\beta_j^2\right)$$

where λ is the ridge parameter, which controls the amount of shrinkage applied to the coefficients and β is the vector of regression coefficients. Larger values of λ lead to more shrinkage and, consequently, more bias but less variance in the estimates. It essentially adds a penalty equal to the square of the magnitude of coefficients to the loss function:

$$\min\left(|Y - X\beta|^2 + \lambda|\beta|^2\right)$$

where λ is the regularization parameter that controls the strength of the penalty. Thus, ridge regression shrinks the coefficients of correlated predictors and reduces their variance. It does not eliminate multicollinearity but minimizes its effects. But the usefulness of this technique is limited by the fact that it introduces bias into the estimates to reduce variance and improve prediction accuracy, making the interpretation of coefficients difficult.

Lasso (Least Absolute Shrinkage and Selection Operator) **Regression** minimizes a similar objective function, but with the $L1$ penalty instead of the $L2$ penalty used in Ridge Regression:

$$min_{\beta} \Sigma\left(y_i - \beta_0 - \Sigma\left(x_i j * \beta_j\right)\right)^2 + \lambda * \Sigma\left(\left|\beta_j\right|\right)$$

$$\Rightarrow \min\left(\left|Y - X\beta\right|^2 + \lambda\left|\beta\right|_1\right)$$

Lasso can shrink some coefficients exactly to zero, thereby performing variable selection and potentially removing multicollinear predictors. But, like ridge regression, it introduces bias and can be challenging to choose the optimal λ.

Elastic Net Regression combines penalties of ridge and LASSO:

$$\min\left(\left|Y - X\beta\right|^2 + \lambda_1\left|\beta\right|^2 + \lambda_2\left|\beta\right|_1\right)$$

This technique balances between ridge and LASSO, shrinking coefficients and performing variable selection but requires tuning of two parameters, which can be computationally intensive.

Prior to applying methods like ridge, lasso, or elastic net, it is often recommended to standardize predictors to have mean 0 and variance 1. This ensures that penalties are applied uniformly across predictors. These techniques can help stabilize the model estimates in the presence of multicollinearity, but they introduce bias into the estimates. Additionally, selecting the appropriate value for the tuning parameter (**λ) can be challenging.**

Principal Component Regression: Principal Component Analysis (PCA) is a technique that transforms the original variables into a new set of uncorrelated variables called principal components. These components are linear combinations of the original variables and account for most of the variance in the data. The steps followed are: First, conduct principal component analysis on the predictors to transform them into a set of linearly uncorrelated components. Then, use these components as predictors in a regression model. Mathematically, if $\boldsymbol{X}$ is the matrix of predictors, PCA transforms $\boldsymbol{X}$ into a new set of variables: $\boldsymbol{Z} = \boldsymbol{XW}$, where $\boldsymbol{W}$ is the matrix of eigenvectors of $\boldsymbol{X'X}$, and then regression is done on $\boldsymbol{X}$ instead of $\boldsymbol{X}$. The first principal component is the linear combination of the original variables that maximizes the variance:

$$var(Y_1) = \sum(eigenvectors_1 * eigenvalues_1)$$

where $eigenvectors_1$ are the eigenvectors corresponding to the largest eigenvalue. Thus, by using principal components (which are uncorrelated) as predictors, multicollinearity is eliminated. Although PCA can help reduce dimensionality and remove multicollinearity, but it can be challenging to interpret the principal components, and this method may not be appropriate when the goal is to understand the effects of the original variables on the response.

Partial Least Squares Regression (PLSR): PLSR is similar to PCR but constructs new predictor variables (latent variables) that account for both the variance in the predictor variables and the covariance between the predictor variables and the response variable. Let X be the matrix of predictor variables, and T be the matrix of latent variables. The PLSR model can be written as:

$$y = \beta_0 + T * \beta + \varepsilon$$

where β is the vector of regression coefficients for the latent variables. PLSR can be computationally intensive, and the choice of the number of latent variables can be subjective.

Variable Selection Methods: These methods involve removing one or more highly correlated predictor variables from the regression model, thereby reducing multicollinearity.

Stepwise Regression: This method involves sequentially adding or removing predictor variables based on their statistical significance in the model. Variables with high multicollinearity are typically removed.

Best Subset Selection: This method involves evaluating all possible combinations of predictor variables and selecting the best subset based on a predetermined criterion, such as adjusted R-squared or Akaike Information Criterion (AIC).

VIF Stepwise Selection: This method involves iteratively removing the variable with the highest VIF until all remaining variables have a VIF below a predetermined threshold. This approach can be effective in reducing multicollinearity, but it might lead to the omission of important variables.

Variable selection methods can lead to biased coefficient estimates and may exclude important variables from the model.

After addressing multicollinearity, it is crucial to use appropriate model selection criteria (e.g., AIC, BIC, cross-validation) to evaluate the performance and predictive accuracy of the model, ensuring that the chosen method effectively balances bias and variance.

Multicollinearity poses challenges but can be managed through various statistical techniques, each with its rationale and limitations. While these methods can help mitigate the effects of multicollinearity, they may also introduce additional complexities or limitations. Researchers should carefully consider the trade-offs and assumptions associated with each method, as well as the specific characteristics of their dataset and research objectives, when deciding on the most appropriate approach. The choice of method depends on the specific context, including the severity of multicollinearity, the goal of the analysis (interpretation vs. prediction), and the characteristics of the data.

4.2.9 Summary

The classical linear regression model is based on a set of assumptions that ensure valid statistical inference. Among these, the non-stochastic assumptions deal specifically with the independent variables, also known as regressors, denoted by the matrix X. These assumptions are called "non-stochastic" because they relate to the fixed and predetermined independent variables, which are not subject to random variation. The main non-stochastic assumptions are:

1. Correct model specification: The model correctly specifies the functional form, includes all relevant variables, and excludes irrelevant ones. Incorrect specification, such as omitting a relevant variable or including an irrelevant one, can lead to biased and inconsistent parameter estimates.
2. Linear in parameters: The model is assumed to be linear in parameters, meaning the relationship between the independent variables and the dependent variable is linear.
3. Fixed and deterministic X: The values of the independent variables in the sample are fixed and known without error, implying no measurement error associated with the independent variables.
4. No endogeneity: There is no two-way causal relationship between the independent variables and the dependent variable. The

independent variables solely influence the dependent variable and are not influenced by the error term.

5. No multicollinearity: The independent variables in X are not perfectly linearly dependent on each other. Perfect collinearity would prevent the unique estimation of the coefficients.
6. Full-rank X: The number of observations is greater than the number of independent variables, and the rank of the X matrix equals the number of independent variables, ensuring no redundant rows or exact linear relationships within the independent variables.

Correct model specification is paramount, ensuring the model accurately represents the underlying relationship by including all relevant variables while excluding irrelevant ones, thereby preventing bias and inconsistency in parameter estimates. The assumption of linearity in parameters implies that the relationship between independent and dependent variables is linear, albeit allowing for transformations of variables to meet this criterion.

The fixed and deterministic X assumption treats the independent variables as constants, assuming no measurement error and eliminating the possibility of serial correlation by not including lagged dependent variables as regressors. No endogeneity assumes a one-way causal relationship from the independent variables to the dependent variable, without any feedback from the latter. This prevents the independent variables from being correlated with the error term, which could introduce bias and inconsistency in the regression coefficients.

No multicollinearity requires that the independent variables are not perfectly linearly dependent on each other, ensuring the model's ability to uniquely estimate coefficients. The full-rank X assumption, necessitating more observations than variables and no redundant linear relationships among variables, supports the model's estimability and interpretability.

Violations of these assumptions, such as autocorrelation and multicollinearity, can lead to biased, inefficient estimates and unreliable statistical inferences. The violation of the assumption of serial independence, also known as autocorrelation, occurs when the error terms in the model are correlated over time or across observations. Autocorrelation can arise due to various reasons, such as omitted variables, model misspecification, or the presence of dynamics or persistence in the dependent variable or error

terms. The consequences of autocorrelation include inefficient estimates, invalid statistical inferences, and, in some cases, biased estimates.

Several tests are available to detect autocorrelation, including the Von Neumann Ratio, Durbin-Watson test, Breusch-Godfrey test, and Ljung-Box test. These tests are designed to identify patterns or dependencies in the error terms, indicating the presence of autocorrelation.

If autocorrelation is detected, various remedies can be employed, such as autoregressive models (AR), moving average models (MA), autoregressive integrated moving average (ARIMA) models, generalized autoregressive conditional heteroskedasticity (GARCH) models, generalized least squares (GLS), the Cochrane-Orcutt procedure, and the Newey-West standard errors.

The assumption of non-multicollinear regressors implies that the independent variables in the regression model should not be perfectly correlated with one another. Multicollinearity can lead to infinitely large standard errors of estimates, indeterminate estimation of regression coefficients, and misleading interpretation of regression coefficients.

Several tests and diagnostics are used to detect multicollinearity, including Frisch's Confluence Analysis, the Farrar-Glauber test, Variance Inflation Factor (VIF), tolerance, condition index, eigenvalues of the correlation matrix, and the correlation matrix itself.

When multicollinearity is detected, various solutions can be employed, such as increasing the sample size, removing correlated variables, creating linear combinations of variables, using regularization techniques (ridge regression, LASSO, elastic net regression), principal component regression, partial least squares regression (PLSR), and variable selection methods (stepwise regression, best subset selection, VIF stepwise selection).

The choice of solution depends on the specific context, including the severity of multicollinearity, the goal of the analysis (interpretation vs. prediction), and the characteristics of the data. Each solution has its rationale, assumptions, and limitations, which should be carefully considered before implementation.

4.2.10 Keywords

Correct Model Specification: Accurately representing the relationship between dependent and independent variables by including all relevant variables and excluding irrelevant ones, to prevent biased estimates.

Linear in Parameters: The model assumes a linear relationship between the dependent variable and the parameters, albeit allowing transformations of the variables themselves.

Fixed and Deterministic X: Independent variables in the model are considered fixed and known with certainty, with no measurement errors, ensuring no serial correlation from using lagged dependent variables as regressors.

No Endogeneity: Independent variables should influence the dependent variable without being influenced by it in return, avoiding correlation with the error term that leads to biased estimates.

No Multicollinearity: Independent variables are not perfectly linearly dependent on each other, enabling the model to uniquely estimate coefficients.

Full-Rank X: More observations than variables, with no redundant or perfectly collinear variables, ensuring the estimability of the model.

Assumption of Serial Independence: Error terms in the regression model are not correlated with each other, ensuring no autocorrelation and maintaining efficiency in estimates.

Consequences of Autocorrelation: Autocorrelation leads to inefficient estimates and unreliable statistical inferences, potentially causing biased estimates in certain scenarios.

Tests for Autocorrelation: Diagnostic tests like Von Neumann Ratio, Durbin-Watson, Breusch-Godfrey, and Ljung-Box are used to detect autocorrelation.

Estimation with Autocorrelation: Solutions to autocorrelation include autoregressive models, moving average models, ARIMA, GARCH, generalized least squares, Cochrane-Orcutt procedure, and Newey-West standard errors.

Consequences of Multicollinearity: Problems caused by multicollinearity, such as indeterminate and inflated estimates of regression coefficients.

Tests for Multicollinearity: Methods like Frisch's Confluence Analysis, Farrar-Glauber Test, Variance Inflation Factor (VIF), Condition Index, eigenvalues of the correlation matrix, and correlation matrix inspection are utilized to detect multicollinearity.

Solutions for Multicollinear Data: Increasing sample size, removing correlated variables, combining variables, regularization techniques (ridge regression and lasso), Principal Component Regression, and Partial Least Squares Regression as remedies.

Von Neumann Ratio Test: Detects first-order autocorrelation by comparing the mean square successive difference to the variance of time series.

Durbin-Watson Test: Widely used test for detecting first-order autocorrelation in regression models, with simple interpretation and critical values.

Breusch-Godfrey Test: Extends the Durbin-Watson test to detect higher-order autocorrelation, relying on an auxiliary regression of residuals on original regressors and their lagged values.

Ljung-Box Test: Investigates the presence of autocorrelation in time series or residuals by testing the randomness of a sequence, suitable for continuous data.

Frisch's Confluence Analysis: Identifies sources of multicollinearity by decomposing regressors into components free from and containing multicollinearity.

Farrar-Glauber Test: Tests the significance of regression coefficients when individual regressors are removed, indicating multicollinearity if coefficients significantly change.

Variance Inflation Factor (VIF) and Tolerance: Quantify how much the variance of an estimated regression coefficient is increased due to multicollinearity, with higher VIF values indicating greater multicollinearity.

Regularization Techniques: Techniques like ridge regression, lasso, and elastic net are used to address multicollinearity by adding a penalty to the loss function, reducing the impact of highly correlated predictors.

Partial Least Squares Regression (PLSR): Constructs new predictor variables that account for the variance in predictors and the covariance

between predictors and the response variable, offering an alternative when multicollinearity is present.

Principal Component Regression (PCR): Utilizes principal component analysis to transform correlated variables into a set of uncorrelated principal components used as predictors in the regression model, effectively addressing multicollinearity.

Condition Index: A diagnostic measure used to assess the severity of multicollinearity based on the singular value decomposition of the design matrix, with higher values indicating stronger multicollinearity.

Eigenvalues of the Correlation Matrix: Inspection of eigenvalues from the correlation matrix of predictors, where smaller eigenvalues suggest the presence of multicollinearity among the variables.

Correlation Matrix Inspection: A simple method to detect multicollinearity by examining the pairwise correlation coefficients among independent variables, looking for high correlation coefficients as indicators of multicollinearity.

Regularization Parameter (λ): In regularization techniques like ridge and lasso regression, λ controls the amount of shrinkage applied to the coefficients, balancing the bias-variance trade-off.

Ridge Regression: A regularization technique that addresses multicollinearity by adding a penalty equal to the square of the magnitude of coefficients, which helps in reducing the variance of coefficient estimates.

Lasso Regression: A regularization technique that, unlike ridge regression, can shrink some coefficients to zero, thus performing variable selection in addition to addressing multicollinearity.

Elastic Net Regression: Combines the penalties of both ridge and lasso regression, making it suitable for situations where there is multicollinearity among predictors, as well as when there is a need for variable selection.

Stepwise Regression and Best Subset Selection: Variable selection methods that involve either sequentially adding or removing variables based on their statistical significance, or evaluating all possible combinations of variables, respectively, to address issues including multicollinearity.

4.2.11 Self-assessment Questions

1. What is the significance of the assumption of fixed and deterministic X in classical linear regression models?
2. Explain how multicollinearity affects the precision and interpretation of regression coefficients.
3. Describe the consequences of autocorrelation in the error terms of a linear regression model.
4. What does the assumption of no endogeneity imply about the relationship between independent variables and the error term?
5. How does correct model specification ensure unbiased and consistent parameter estimates in linear regression?
6. Discuss the implications of violating the linear in parameters assumption for a regression model.
7. Why is the full-rank X assumption crucial for the estimability of regression coefficients?
8. How can the presence of multicollinearity be detected in a regression model?
9. What remedies are available for dealing with autocorrelation in time-series data analysis?
10. Explain the importance of the assumption of serial independence in the context of regression errors.
11. Given the following data on Y (dependent variable) and X (independent variable), calculate the simple linear regression coefficients (β_0 and β_1) using the OLS method.

Y	X
2	1
3	2
5	3
7	4
9	5

12. Using the residuals from Question 11, test for autocorrelation using the Durbin-Watson statistic.

13. Consider a dataset with two independent variables (X_1 and X_2) showing signs of multicollinearity. Calculate the Variance Inflation Factor (VIF) for X_1.

X_1	X_2	Y
1	2	5
2	4	10
3	6	15
4	8	20

14. Given the following error terms from a regression model, calculate the first-order autocorrelation coefficient (ρ) using the formula.

t	e_t
1	0
2	1
3	0.5
4	–0.5
5	–1

15. For the dataset given in Question 13, calculate the coefficients using Ridge Regression with λ=0.5. Assume the intercept is zero for simplification.
16. Using the dataset from Question 11, compute the adjusted R-squared value for the simple linear regression model.
17. Given the following data, perform a Breusch-Godfrey Test for up to 2nd order autocorrelation. Use a significance level of 5%.

t	Residual (e_t)
1	–0.2
2	0.1
3	–0.1
4	0.2
5	–0.1

18. For the dataset in Question 13, calculate the Elastic Net Regression coefficients with λ1=0.1 and λ2=0.1, assuming the intercept is zero for simplification.
19. Using the given hypothetical data for a time series, calculate the Ljung-Box Q statistic for up to 3 lags. Assess the presence of autocorrelation.

t	e_t
1	0.3
2	0.1
3	–0.2
4	0.4
5	–0.3

20. Given the following data on three independent variables (X_1, X_2, X_3) and Y, perform a principal component analysis and calculate the first principal component scores. Use the scores as predictors in a simple linear regression model to estimate Y.

X_1	X_2	X_3	Y
10	20	5	30
20	40	10	60
30	60	15	90
40	80	20	120

4.2.12 References

1. **Applied Econometric Time Series** by Walter Enders. Focused on time series econometrics, this book is excellent for understanding topics like autocorrelation and forecasting models. It's approachable for undergraduates, especially those interested in the application of econometrics in financial and macroeconomic data.
2. **Using Econometrics: A Practical Guide** by A.H. Studenmund. Aimed at beginners, this book offers a practical, step-by-step approach to learning econometrics, with a strong emphasis on

understanding and applying econometric concepts rather than on mathematical derivations.

3. **Introduction to Econometrics** by James H. Stock and Mark W. Watson. This text is known for its intuitive approach and focuses on the practical implementation of econometric methods, making it an excellent resource for understanding the basics of regression analysis and its assumptions.
4. **Applied Econometrics with R** by Christian Kleiber and Achim Zeileis. For students interested in practical application and programming, this book introduces econometric models, including linear regression and its assumptions, through the R statistical programming language.
5. **A Guide to Modern Econometrics** by Marno Verbeek. This book offers an accessible yet thorough overview of econometric theory and practice, including discussions on the challenges and solutions related to regression analysis assumptions.

UNIT – V : Regression on Dummy Independent Variables

Lesson 5.1 – The Nature of Dummy Variables

Structure

5.1.1 Introduction

Econometric models often rely on numerical data, but many interesting economic questions involve categories (e.g., gender, race, industry type). Categorical data represents qualities or classifications that do not have a natural numerical scale. For instance, we cannot meaningfully add or subtract values like "high school diploma" or "unemployed." Categorical variables, such as gender, race, educational level, or geographical regions, cannot be directly included in a linear regression model because they lack a natural numerical scale. Dummy variables provide a way to encode these categorical variables into a format that can be used in regression analyses. By including dummy variables in a regression model, researchers can quantify the impact of these categorical factors on the outcome variable of interest.

Dummy variables act as a bridge between categorical data and regression models. They convert categories into simple 0s and 1s, allowing researchers to analyze how these categories influence a continuous outcome. In essence, these variables (also called binary, indicator, or categorical variables) translate qualitative information into a format usable by regression analysis. This lets researchers incorporate non-numerical factors and assess their impact on the variable being predicted.

Coefficients of dummy variables are straightforward to interpret, representing the average difference in the dependent variable for the category relative to the reference category. By including dummy variables for potential confounders, researchers can more accurately estimate the effect of the variables of interest.

A dummy variable takes the value of 1 or 0 to indicate the presence or absence of some categorical effect that may be expected to shift the outcome. For example, suppose we are studying the effect of education level on income. We have a category for "high school diploma." To include this in a regression, we create a dummy variable. This variable will take a value of 1 if someone has a high school diploma and 0 otherwise. The coefficient of this dummy variable in the regression analysis will tell us how much, on average, having a high school diploma (compared to the reference category, typically those without a diploma) affects income. Consider another example: in studying wage determinants, a dummy variable for gender might be created where 1 represents female and 0 represents male. This allows the model to account for wage differences that are attributable to gender.

In econometric time series analysis, dummy variables may be used to indicate the occurrence of wars, major strikes, or other significant events). In panel data analyses, dummy variables are used to control for unobserved heterogeneity across entities (e.g., individuals, firms, countries) by including a dummy for each entity. This fixed effects approach helps isolate the effect of variables of interest from the influence of time-invariant characteristics.

Dummy variables can be used to test for interaction effects between categorical and continuous variables. For example, by creating an interaction term between a dummy variable (e.g., receiving training) and a continuous variable (e.g., hours of training), analysts can examine whether the impact of training on productivity varies with the amount of training.

Dummy variables are essential for modeling seasonal patterns or trends over time. Dummy variables can be added to represent each of the possibly many seasonal periods contained in data, such as hourly or weekly variations in traffic flow. For instance, monthly data analyses might include 11 dummy variables to capture monthly effects (with one month omitted as a reference), allowing the model to adjust for seasonal variations in the data.

Dummy variables can help control for confounding factors and improve the validity of results). For instance, when building a model to explain income in a cross-section of the population, a dummy variable for gender could be included to test the hypothesis that men have a higher starting salary and faster trajectory than women.

Dummy variables can also be used to represent subgroups of the sample in a study—often to distinguish different treatment groups. In the simplest case, a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables enable the use of a single regression equation to represent multiple groups, eliminating the need to write out separate equation models for each subgroup. Dummy variables can account for group-specific effects that might otherwise cloud your results. In a treatment effects model, a dummy variable can represent the effect of both time (before and after treatment) and group membership (treatment or control group).

Thus, dummy variables are a powerful tool in econometric analyses for representing categorical data, indicating key events, controlling for confounding factors, modeling seasonal and treatment effects, and more. They allow for more nuanced models that capture the influence of categorical variables. However, they should be used judiciously to ensure valid and generalizable models.

5.1.2 Dummy Variables and Intercept Parameters

Consider the data of Gross Domestic Consumption and National Income of India for the period 1981-81 to 2022-23 given in Table–1. This period saw three major crises: the economic crisis of 1990-1991, the worldwide recession 2008-09, and the COVID–19 lockdowns 2020-21. Apart from these, the Indian economy also suffered from the shock of demonetization towards the end of 2016 and the introduction of Goods and Services Tax (GST) a year later. While GST implementation impacted

the small and medium enterprises (MSMEs) mainly, the others–economic crises and demonetization–hurt the consumers equally or even more than they did the producers.

Now, if we were to model the relationship between national income and gross domestic consumption for the period 1981-82 to 2022-23, we must take into account these periods of crises where the relationship between income and consumption was severely impacted upon by external forces. Not accounting for these periods will lead to unreliable estimates. In the absence of any major episodes, the relationship between gross domestic consumption and national income could be represented by a simple linear regression model, such as:

$$C_t = \beta_0 + \beta_1 Y_t + \grave{o}_t \tag{5.1}$$

where $Y_t = \text{national income}$ and $C_t = \text{gross domestic consumption}$

But this equation does not take into account the impact of the crises on the domestic consumption. The economic crises are expected to shift this consumption function *downwards* during the crisis periods. How do we account for that?

Dummy variables are a way to incorporate the effect of these economic episodes into the relationship between gross domestic consumption and national income. Essentially, we create a dummy variable, let's say *D*, and assign it a value of 1 if the gross domestic consumption and national income data refer to an year of economic crisis and a value of 0 if the data corresponds to an episode-free year. Thus, the dummy variable acts like a switch which is turned 'on' (value = 1) during crisis years and turned 'off' (value = 0) during non-crisis years. Thus:

$$D = \begin{cases} 1 \text{ crisis years} \\ 0 \text{ normal years} \end{cases}$$

Having created a dummy variable to represent the crisis years, we must now decide how did the crises impact the domestic consumption. This is the most crucial step with only theory to guide. If we theorize that the economic crises impacted the level of *autonomous* consumption in the economy, then we essentially mean that the intercept, β_0, in equation 5.1 does not remain constant for all values in the sample years. We model this mathematically by modifying equation 5.1 as:

$$C_t = \beta_{0t} + \beta_1 Y_t + \grave{o}_t \tag{5.2}$$

Notice that the intercept parameter now has a time subscript attached to it indicating the fact that it is no longer a constant rather it changes from one observation to another. Thus, we model the intercept incorporating the dummy variable as:

$$\beta_{0t} = \beta_0 + \delta D_t \tag{5.3}$$

Since we have theorized that the effect of crises is on the level of autonomous consumption, the coefficient of the dummy variable in equation 5.3, represents the reduction in gross domestic consumption during the economic episode years and is expected to have a negative sign. We incorporate this assumption about the dummy variable to expand our linear relationship model as:

$$C_t = \beta_0 + \delta D_t + \beta_1 Y_t + \text{ö}_t \quad t = 1981, \ldots, 2023 \tag{5.4}$$

Table 1: Domestic Consumption and National Income of India (Source: Handbook of Indian Economy, RBI)

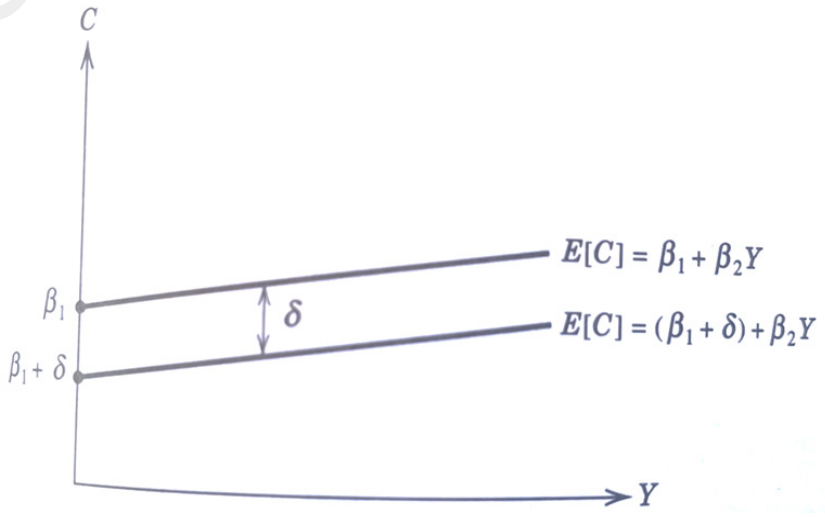
Year	National Income (Rs. Billion)	Gross Domestic Consumption (Rs. Billion)	Year	National Income (Rs. Billion)	Gross Domestic Consumption (Rs. Billion)
1981-82	1331.95	684.58	2002-03	15465.04	7671.3
1982-83	1498.13	728.83	2003-04	17462.7	8722.13
1983-84	1766.84	872.22	2004-05	19802.26	9994.76
1984-85	1942.88	962.11	2005-06	22649.38	11487.77
1985-86	2187.35	1092.36	2006-07	26242.95	13464.75
1986-87	2507.21	1228.24	2007-08	31127.15	15977.93
1987-88	2891.02	1392.25	2008-09	36907.09	19182.54
1988-89	3507.91	1703.02	2009-10	44706.5	23284.63
1989-90	4090.26	1970.18	2010-11	53031.49	27693.06
1990-91	4598.74	2199.5	2011-12	61759.31	32440.52
1991-92	5020.13	2406.13	2012-13	67599.35	35756.22
1992-93	5448.59	2651.3	2013-14	73782.13	38845.69

Year	National Income (Rs. Billion)	Gross Domestic Consumption (Rs. Billion)	Year	National Income (Rs. Billion)	Gross Domestic Consumption (Rs. Billion)
1993-94	6144.45	2938.54	2014-15	80709.89	42607.96
1994-95	6954.98	3369.62	2015-16	88171.14	46639.62
1995-96	8071.98	3919.47	2016-17	96412.28	51089.12
1996-97	9134.43	4436.55	2017-18	104756.92	55858.28
1997-98	10142.92	4860.6	2018-19	114435.85	60850.47
1998-99	11350.89	5510.28	2019-20	119363.78	63425.47
1999-00	12413.05	6158	2020-21	118318.33	62517.59
2000-01	13277.45	6623.97	2021-22	130225.85	69231.3
2001-02	14315.04	7123.41	2022-23	142424.47	75776.3

The econometric model we have hypothesized changes as the value of the dummy variable changes during the crisis and non-crisis periods:

$$C_t = \begin{cases} \beta_0 + \beta_1 Y_t & \text{when } D_t = 0 \\ (\beta_0 + \delta) + \beta_1 Y_t & \text{when } D_t = 1 \end{cases}$$

Thus, during the crisis years, the intercept parameter, β_0, is *reduced* by the amount δ as it is expected to carry a negative sign. This can be presented graphically as in Figure–1 below:



The dummy variable, D_t, is called the intercept dummy variable as it captures a shift in the intercept of a linear relationship modeled by a regression equation. The dummy variable, in this case, has the effect of pulling the consumption line *downwards* as it is expected to have a negative value.

If the error terms, $\dot{o}_t$, in equation 5.4 follow the assumptions of the classical linear regression model, all the parameters, including that of the dummy variable, can be estimated using the ordinary least squares (or any other, such as the maximum likelihood) estimation procedure. The dummy variable is treated as any other explanatory variable in the equation and the estimation is not affected by the fact that it consists of only zeros and ones.

The estimated value of the parameter δ can be checked for statistical significance using the normally employed test of hypothesis; if the null hypothesis of $\delta = 0$ is not rejected, then we can conclude that the economic crises had no effect on the level of autonomous consumption.

For the data in Table–1, the economic crisis years are judged to be 1990-91 to 1992-93, 2008-09 to 2010-11, 2016-17 to 2017-18, and 2020-21 to 2021-22; the dummy variable is assigned a value of 1 for these years, 0 for the rest of the years. Running the ordinary least squares procedure result in the following estimates:

$$\hat{C}_t = 344.72 - 137.58D_t + 0.53Y_t \tag{5.5}$$

The standard errors of these parameter estimates are as follows:

$$SE_{b_0} = 0.001$$
$$SE_{\delta} = 7.596$$
$$SE_{b_1} = 41.073$$

The adjusted coefficient of determination turns out to be: $R^2 = 0.999$ indicating a very high explanatory power of the hypothesized regression model. The omnibus test of significance of all the parameter estimates together being non-zero yields: $F = 28.90$, significant at $\alpha = 0.01$.

As expected, the ordinary least squares estimate of the dummy variable parameter, δ, is negative. Thus, on an average, the level of autonomous domestic consumption decreased by 137.58 billion rupees during the crisis years. This drop in the level of domestic consumption is the manifestation

of structural changes in the economy during the crisis years. The regression equation also shows that the marginal propensity to consume, estimated by b_1, is very low at just 53% of the marginal income.

5.1.3 Dummy Variables and Slope Parameters

The effect of a qualitative variable has to be guided by the theory. Suppose we have an *a priori* reason to argue that the level of autonomous consumption was not affected by the economic shocks, rather the marginal propensity to consume–represented by the coefficient of the income variable–was affected. To model this relationship we modify equation 5.1 as:

$$C_t = \beta_0 + \beta_{1t} Y_t + \grave{o}_t \tag{5.6}$$

Where the addition of the subscript t to the slope parameter indicates that it is no longer the same for all years and changes from one period to another. As earlier, a dummy variable is created to incorporate the fact that the slope parameter changes with cases. The slope parameter can now be modeled as:

$$\beta_{1t} = \beta_1 + \gamma D_t \tag{5.7}$$

We assume, for simplicity of exposition, that the intercept parameter is constant. The coefficient γ of the dummy variable D_t, represents the change in the marginal propensity to consume during the crisis and non-crisis years in the sample.

Incorporating the effect of economic crises in the regression model implies substituting equation 5.7 into equation 5.6 to obtain:

$$C_t = \beta_0 + \left(\beta_1 + \gamma D_t\right) Y_t + \grave{o}_t \tag{5.8}$$

Or, equivalently as:

$$C_t = \beta_0 + \beta_1 Y_t + \gamma D_t Y_t + \grave{o}_t \quad t = 1981, \ldots, 2023 \tag{5.9}$$

This representation essentially means that modeling the assumption of time varying marginal propensity to consume adds a composite variable to the regression relationship: $D_t Y_t$, the product of the dummy variable and national income. This composite is called an interaction variable and it captures the interaction effect of economic crises and national income of domestic consumption. This interaction variable is absent during the normal years as the dummy variable takes the value of zero and equals the

national income during the crisis years when the dummy variable takes the value of one. Thus, the interaction variable changes the consumption function during the periods of economic crises:

$$C_t = \begin{cases} \beta_0 + \beta_1 Y_t & \text{when } D_t = 0 \\ \beta_0 + (\beta_1 + \gamma) Y_t & \text{when } D_t = 1 \end{cases}$$

Thus, during normal years, the marginal propensity to consume is β_1 and this changes during episodes of economic crises to $(\beta_1 + \gamma)$; as earlier, we expect γ to be negative. Provided that the error terms in equation 5.9 follow the assumptions of the classical linear model, the parameters of the equation, $(\beta_0, \beta_1, \gamma)$, can be estimated using the ordinary least squares procedure. The hypothesis that the episodes of economic distress had no effect on the marginal propensity to consume can be tested the usual way for the null hypothesis: $H_0 : \gamma = 0$

5.1.4 The Intercept Dummy and the Interaction Variable

If we relax the somewhat artificial assumption of constant intercept and accept the more likely scenario that the economic shocks affected both the level of autonomous consumption and the marginal propensity to consume, then we need to incorporate both the shift in the intercept and time varying slope into the econometric mode. This can be achieved by putting together equations 5.4 and 5.9 to get:

$$C_t = \beta_0 + \delta D_t + \beta_1 Y_t + \gamma (D_t Y_t) + \varepsilon_t \quad t = 1981, \ldots, 2023 \qquad (5.10)$$

Where, as before, δ is the parameter of the shift in the intercept and γ is the parameter of the interaction variable. Notice that, there is only one dummy variable, D_t, representing the presence or absence of the economic shocks. Provided that the error terms of equation 5.10 follow the assumptions of the classical linear regression model, the parameters of the equation can be estimated, as before, using the ordinary least squares procedure and tested for statistical significance.

5.1.5 Dummy Variable for the Dependent Variable

So far, we have considered cases where the outcome variable was continuous and only the explanatory variables were categorical. But there may be cases where the outcome variable itself is categorical. Consider, for example, the very practical problem of predicting someone as potential buyer of a luxury smartphone. Suppose that the ownership of a luxury

smartphone of a certain brand is found to be heavily influenced by the income earned by the individual and the profession to which the individual belongs. This relationship is then modeled as:

$$C_i = \beta_0 + \beta_1 Y_i + \beta_2 W_i + \grave{o}_i \tag{5.11}$$

Where:

C_i = individual owns a luxury smartphone

Y_i = income of the individual

W_i = profession of the individual

Here the ownership of the luxury smartphone is clearly a dichotomous variable—an individual owns it or does not own it. It can be represented as a dummy variable that takes the value 1 if the individual owns a luxury smartphone and the value 0 if the he does not own a luxury smartphone. That is:

$$C_i = \begin{cases} 0 \text{ if does not own a luxury smartphone} \\ 1 \text{ if does own a luxury smartphone} \end{cases}$$

In case of dichotomous outcome variables, the error terms are *not* homoscedastic. And, thus the ordinary least squares procedure is not appropriate to estimate the model parameters. Such cases are estimated with a modified set of variables and the estimation techniques fall under the category of binary logistic regression models, namely the logit and the probit models. Discussion of these models is beyond the scope of this book; they can be learned in advanced courses of econometrics.

5.1.6 Using Dummy Variables in Case of More than Two Classes

The dummy variables that we have studied so far have only two states: a value of 0 or 1. But in the example considered in the last section, the explanatory variable 'profession' can, rather will, have more than two values, e.g., investment banker, industrialist, doctor, academician, politician, etc. Creating a dummy variable for 'profession' needs some consideration.

Consider the case where we want to model the consumption of caviar by customers who buy it from a upmarket superstore so that we are able to predict the amount of caviar that will be bought by a particular customer who walks into the superstore. In a cross-sectional study of this kind the prices of the goods concerned and of all its substitutes and compliments are a given and therefore are not included in the econometric model.

Suppose that a previous focus group study that found that the factors that affect the caviar consumption are mainly: income, age, gender, and profession. While the age and income of the individual are continuous variables, gender and profession are not. Simple categorical variables like gender take on only two values and it is easy to create dummy variables for them. For example, the gender can be incorporated as a dummy variable with two values:

$$G = \begin{cases} 0 \text{ if male} \\ 1 \text{ if female} \end{cases}$$

But for profession, which has more than two states, dummy variables need to be defined for *each* of the possible states. Suppose that the profession has only four values: banker, politician, industrialist, and real estate, then we may create the following dummy variables:

$$W_{Bank} = \begin{cases} 0 & \text{if the customer is not a banker} \\ 1 & \text{if the customer is a banker} \end{cases}$$

$$W_{Politician} = \begin{cases} 0 & \text{if the customer is not a politician} \\ 1 & \text{if the customer is a politician} \end{cases}$$

$$W_{Industrialist} = \begin{cases} 0 & \text{if the customer is not an industrialist} \\ 1 & \text{if the customer is an industrialist} \end{cases}$$

Notice that, for four states of the variable profession, we have created only three dummy variables—one less than the number of states. The omitted state, it can be any state decided arbitrarily, is taken as the *base* case or the *reference* level.

5.1.7 The Dummy Variable Trap

The reason for creating one less dummy variable is that if we create for all states that a particular variables takes, then we will have a case of *exact* multicollinearity. If we create dummy variables for all the states, they will add up to one making them linearly dependent on the intercept variable. That is, the design matrix of X will have an extra variable apart from the intercept that will have all its values as 1, thus rendering two columns of the design matrix having the exact same values and violating the full rank order condition for the estimation of the parameters. This tendency to create dummy variables for all states of a categorical variable is called falling into a 'dummy variable trap'—an eminently avoidable trap.

Let us consider one of the most common use of dummy variables—removing seasonality in a timeseries data. Many timeseries data exhibit seasonal variations due to various factors–for example, air conditioner sales are particularly high during the summers, raincoats during the rainy season, etc. Similarly, sales of consumer electronics and clothing is higher during the Durga Puja and Deepavali festivals. These variations in the data are called seasonality. They can be easily modeled using dummy variables.

Suppose we want to model retail sales of air conditioners in Northern India. They exhibit high volumes during the months of April-May-June and large drops during the month of October-November-December. We can create a dummy variable representing the quarters of a year and incorporate them into the regression model as follows:

$$S_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \delta Q_{1t} + \gamma Q_{2t} + \eta Q_{3t} + \dot{o}_t \qquad (5.12)$$

Where:

$$S_t = \text{quarterly sales}$$
$$X_{1t}, X_{2t}, \ldots, X_{kt} = \text{explanatory factors}$$
$$Q_{1t}, Q_{2t}, Q_{3t} = \text{quarterly dummy variables}$$

These dummy variables take the values 0 and 1 as follows:

$$Q_{1t} = \begin{cases} 1 & \text{in the first quarter} \\ 0 & \text{in all other quarters} \end{cases}$$

$$Q_{2t} = \begin{cases} 1 & \text{in the second quarter} \\ 0 & \text{in all other quarters} \end{cases}$$

$$Q_{3t} = \begin{cases} 1 & \text{in the third quarter} \\ 0 & \text{in all other quarters} \end{cases}$$

We created only 3 dummy variables for 4 quarters. The ordinary least squares estimates of $\delta, \gamma, \text{and } \eta$ will give the seasonal effect of the first, second, and third quarters. In the fourth quarter of the year all other dummy variables (Q_{1t}, Q_{2t}, Q_{3t}) are zero; the seasonal effect of this quarter is estimated by the intercept parameter, β_0.

5.1.8 Using Dummy Variables with Two Qualitative Variables

Lets go back to the caviar consumption example in Section 5.1.6. The basic econometric model that relates caviar consumption to the level of income would be:

$$C_i = \beta_0 + \beta_1 Y_i \quad i = 1, 2, \ldots, N \tag{5.13}$$

To keep the exposition simple, let us assume that the only qualitative variables that affect the consumption of caviar by an individual are gender and the type of profession the person is engaged in; also, assume that these qualitative factors only affect the level of autonomous consumption of caviar. Since the autonomous consumption is measured by the intercept, we can incorporate the qualitative characteristics of the individual consumer by modifying the base econometric model as:

$$C_i = \beta_{0i} + \beta_1 Y_i \quad i = 1, 2, \ldots, N \tag{5.14}$$

The addition of the subscript 'i' to the intercept parameter denotes that it is no longer constant but varies from individual to individual. The effect of the qualitative variables on the intercept can be incorporated as:

$$\beta_{0i} = \beta_0 + \delta G_i + \gamma_1 W_{1i} + \gamma_2 W_{2i} + \gamma_3 W_{3i} \tag{5.15}$$

Where:

$$G_i = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

$$W_{1i} = \begin{cases} 0 & \text{if the customer is not a banker} \\ 1 & \text{if the customer is a banker} \end{cases}$$

$$W_{2i} = \begin{cases} 0 & \text{if the customer is not a politician} \\ 1 & \text{if the customer is a politician} \end{cases}$$

$$W_{3i} = \begin{cases} 0 & \text{if the customer is not an industrialist} \\ 1 & \text{if the customer is an industrialist} \end{cases}$$

And $\delta, \gamma_1, \gamma_2, \gamma_3$ are the coefficients of the dummy variables of gender and profession expressing the effect of these qualitative variables on the consumption of caviar. Incorporating equation 5.15 in equation 5.14 will give the model to be estimated:

$$C_i = \beta_0 + \delta G_i + \gamma_1 W_{1i} + \gamma_2 W_{2i} + \gamma_3 W_{3i} + \beta_1 Y_i \quad i = 1, 2, \ldots, N \tag{5.16}$$

In the above equation the reference or the base case is male real estate agent. The complete enumeration of the effects of different combinations of qualitative variables is as follows:

$$C_i = \beta_0 + \beta_1 Y_i \quad \text{male real estate agent} \quad (5.17a)$$
$$C_i = \beta_0 + \gamma_1 + \beta_1 Y_i \quad \text{male banker} \quad (5.17b)$$
$$C_i = \beta_0 + \gamma_2 + \beta_1 Y_i \quad \text{male politician} \quad (5.17c)$$
$$C_i = \beta_0 + \gamma_3 + \beta_1 Y_i \quad \text{male industrialist} \quad (5.17d)$$
$$C_i = \beta_0 + \delta + \beta_1 Y_i \quad \text{female real estate agent} \quad (5.17e)$$
$$C_i = \beta_0 + \delta + \gamma_1 + \beta_1 Y_i \quad \text{female banker} \quad (5.17f)$$
$$C_i = \beta_0 + \delta + \gamma_2 + \beta_1 Y_i \quad \text{female politician} \quad (5.17g)$$
$$C_i = \beta_0 + \delta + \gamma_3 + \beta_1 Y_i \quad \text{female industrialist} \quad (5.17h)$$

These equations can be used to answer questions such as: Do men and women differ significantly in their caviar consumption habits? Does the profession to which one belongs has any significant effect on the level of caviar consumption?

Provided that equation 5.16 follows the assumptions of the classical linear regression model, the ordinary least squares estimates of the parameters can be obtained and tested in the usual way for their statistical significance. For example, the difference in caviar consumption habits of men and women can be tested by examining $H_0 : \delta = 0$ using the Student's t-test. To test whether the caviar consumption level of a banker is the same as that of a politician, we can test $H_0 : \gamma_1 - \gamma_2 = 0$ using a t- or an F-test.

5.1.9 Using Dummy Variables with Continuous Quantitative Variables

Suppose we are investigating whether having high blood pressure is a risk factor for cardiovascular events, controlling for age and cholesterol levels. For this study we gathered data from a sample of patients presented in Table–2.

Table 2: Cardiovascular Event and Blood Pressure

ID	Systolic BP	Diastolic BP	Age	Cholesterol	Cardiovascular Event
1	142	99	30	196	FALSE
2	147	100	71	289	TRUE

ID	Systolic BP	Diastolic BP	Age	Cholesterol	Cardiovascular Event
3	172	97	52	245	FALSE
4	115	81	73	186	TRUE
5	142	95	53	235	FALSE
6	131	64	52	286	FALSE
7	103	69	53	203	TRUE
8	175	63	34	212	FALSE
9	147	83	69	270	TRUE
10	126	91	73	244	TRUE
11	142	61	41	250	FALSE
12	167	75	61	197	FALSE
13	103	97	59	155	FALSE
14	137	84	67	176	FALSE
15	129	70	72	292	TRUE
16	108	103	32	196	FALSE
17	158	84	53	292	TRUE
18	146	103	51	202	FALSE
19	153	99	56	208	FALSE
20	123	102	59	270	FALSE
21	122	69	36	277	FALSE
22	179	105	57	218	TRUE
23	171	63	34	251	FALSE
24	118	107	52	214	TRUE
25	130	84	58	265	FALSE
26	110	61	32	166	TRUE
27	107	64	67	265	TRUE

ID	Systolic BP	Diastolic BP	Age	Cholesterol	Cardiovascular Event
28	125	67	56	200	TRUE
29	137	103	61	178	TRUE
30	157	103	34	242	FALSE
31	116	103	41	290	TRUE
32	107	93	51	281	FALSE
33	147	102	64	222	FALSE
34	125	94	71	236	TRUE
35	109	83	70	201	TRUE
36	116	76	46	262	FALSE
37	131	78	46	170	TRUE
38	101	66	60	189	TRUE
39	105	67	56	166	FALSE
40	174	70	53	216	FALSE

While the dependent variable (the individual experienced a cardiovascular event within the 5-year follow-up period) is dichotomous (True/False), all the explanatory variables are continuous. To classify as high or low blood pressure, we need to convert the blood pressure data into dummy variable.

Based on medical guidelines, we define high blood pressure as systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg. We then create a dummy variable *BP* that takes values 0 and 1 depending on the systolic and diastolic blood pressure reading of patients:

$$BP_i = \begin{cases} 1 & \text{if systolic} \geq 140 \text{ or diastolic} \geq 90 \\ 0 & \text{if systolic} < 140 \text{ and diastolic} < 90 \end{cases}$$

We can now model the relationship between experiencing a cardiovascular event (e.g., heart attack) and age, cholesterol level and high blood pressure. Assuming that having high blood pressure increases the baseline chances of experiencing a cardiovascular event, we define the intercept term of the regression model as:

$$\beta_{0i} = \beta_0 + \delta BP_i$$

We are essentially postulating that the baseline chances of experiencing a cardiovascular event is a patient varying variable; for patients not suffering from high blood pressure, the baseline chance is the same as in the general population, but for patients suffering from high blood pressure, the chances increase by δ, which is expected to carry a positive sign. Therefore, the regression model can be framed as:

$$CVE_i = \beta_0 + \delta BP_i + \beta_1 A_i + \beta_2 C_i + \varepsilon_i \tag{5.18}$$

Where:

CVE_i = the individual experiences a cardiovascular event

A_i = age of the individual

C_i = cholestrol level of the individual

If we have grounds for assuming that blood pressure also affects the accumulation of cholesterol in human body, we can further modify the regression model to incorporate an interaction variable $\left(BP_iC_i\right)$ that will measure how high blood pressure and cholesterol level interact to affect the chances of experiencing a cardiovascular event. In this case, the regression equation can be modified as:

$$CVE_i = \beta_0 + \delta BP_i + \beta_1 A_i + \beta_2 C_i + \gamma\left(BP_iC_i\right) + \varepsilon_i \tag{5.19}$$

But studies have shown that as age increases the blood pressure also increases. This means that there may be an interaction effect of blood pressure and age on the likelihood of experiencing a cardiovascular event. If so, we can modify our equation again to account for this new fact:

$$CVE_i = \beta_0 + \delta BP_i + \beta_1 A_i + \eta\left(BP_iA_i\right) + \beta_2 C_i + \gamma\left(BP_iC_i\right) + \varepsilon_i \qquad i = 1,\ldots,N \tag{5.20}$$

If there is no interaction effect then $\eta, \gamma = 0$ and the equation 5.20 will reduce to equation 5.18 proposed earlier. An *F*-test can be used to check whether all parameter estimates are nonzero simultaneously or not.

An alternative to use is the Binary Logistic Regression, especially since the outcome variable is dichotomous. In that case the logistic regression model would look like:

$$CVE_i = \beta_0 + \beta_1 A_i + \beta_2 S_i + \beta_3 D_i + \beta_4 C_i + \omega_i \tag{5.21}$$

Where:

S_i = systolic blood pressure

D_i = diastolic blood pressure

But using the dummy variable approach has its advantage: if the relationship between blood pressure and cardiovascular risk is not strictly linear, the dummy variable might better capture this threshold effect. Also, for some audiences, the results using the dummy variable may be easier to communicate in the context of established high blood pressure guidelines.

5.1.10 Dummy Variables and the Chow Test

An alternative to using dummy variables to find out structural changes is to use the *F*-test suggested by G.C. Chow and hence also known as the Chow Test. To illustrate, suppose we have data from two separate samples with unequal number of elements but the same variables. Then we can use the two samples individually to generate two separate estimates of the same relationship between the outcome and the predictor variables. We can then check whether the relationship changes from one sample to the other which essentially means that we want to check whether the parameter estimates are equal or not.

Consider the following data on domestic consumption and national income (in billions of rupees) for two different, yet contiguous, time periods of an economy:

Table 3: National Income and Domestic Consumption

Year	National Income	Domestic Consumption	Year	National Income	Domestic Consumption
1950	22480	12616	1960	27584	16506
1951	21987	14858	1961	24811	16741
1952	22426	12455	1962	29717	17603
1953	20996	13894	1963	26216	19858
1954	21315	13622	1964	25955	18495
1955	22299	13718	1965	22848	19887
1956	18616	13714	1966	24119	15524

Year	National Income	Domestic Consumption	Year	National Income	Domestic Consumption
1957	21030	14032	1967	28924	18037
1958	18905	12931	1968	26738	19764
1959	22364	13669	1969	28310	16800

The consumption function for the two separate time periods will have the same variables and the same structural relationship between the outcome and predictor variables but will be estimated using two separate samples of data. Thus, the two regression equations can be formulated as:

$$C_i^A = \beta_0^A + \beta_1^A Y_i^A \tag{5.22}$$

$$C_i^B = \beta_0^B + \beta_1^B Y_i^B \tag{5.23}$$

Formulated in this way, we are interested in knowing whether these two relationships (their estimates, actually) differ in a statistically significant way? This boils down to testing the null hypotheses bout the intercept and the slopes, namely:

$$H_0 : \beta_0^A = \beta_0^B \quad \text{and} \quad H_0 : \beta_1^A = \beta_0^B$$

The intercept null hypothesis asks examines whether the consumption function shifted over the two time periods. The slope null hypothesis tests whether the marginal propensity to consume changed over the two time periods.

To answer these questions, we first, pool together all the observations from both the samples to form a single dataset and then estimate the following relationship:

$$C_{pooled} = \gamma_0 + \gamma_1 Y_{pooled} \tag{5.24}$$

Using the parameter estimates, we calculate the predicted values of the outcome variable (domestic consumption), and then estimate the unexplained variance of the model in equation 5.24:

$$\sum e_{pooled}^2 = \sum c_{pooled}^2 - \sum \hat{c}_{pooled}^2 \tag{5.25}$$

If the two samples contained n_1 and n_2 elements and the number of parameters (intercept and the slopes) is K, then this sum of unexplained variances will have $(n_1 + n_2 - K)$ degrees of freedom. A similar exercise

performed on the two samples separately will yield their respective sum of unexplained variances:

$$\sum e_A^2 = \sum c_A - \sum \hat{c}_A^2 \qquad df = (n_1 - K) \tag{5.26}$$

$$\sum e_B^2 = \sum c_B - \sum \hat{c}_B^2 \qquad df = (n_2 - K) \tag{5.27}$$

We then add the two separate sum of unexplained variances to get the total unexplained variance with $(n_1 - K) + (n_2 - K) = (n_1 + n_2 - 2K)$ degrees of freedom:

$$\sum e_A^2 + \sum e_B^2 \tag{5.28}$$

Next, subtract this from the pooled sum of unexplained variances to get, with $(n_1 + n_2 - K) - (n_1 + n_2 - 2K) = K$ degrees of freedom:

$$\sum e_{pooled}^2 - \left(\sum e_A^2 + \sum e_B^2\right) \tag{5.29}$$

Then to perform the omnibus check that all parameter differences are nonzero, we calculate the following *F* ratio:

$$F^* = \frac{\dfrac{\sum e_{pooled}^2 - \left(\sum e_A^2 + \sum e_B^2\right)}{K}}{\dfrac{\sum e_A^2 + \sum e_B^2}{(n_1 + n_2 - 2K)}} \tag{5.30}$$

At 5% level of significance ($\alpha = 0.05$), if $F^* > F_{0.05}$, we reject the null hypothesis, $H_0 : \beta_i^A = \beta_i^B$ and conclude that the two sets of parameter estimates, and hence the two functions, differ from each other in a statistically significant way. In other words, not only did the consumption function shift from period A to period B but also the marginal propensity to consume changed between the two time periods.

The disadvantage of the Chow Test is that since it is an omnibus test, it does not tell us the cause of the structural change. It can only test for a structural break but cannot pinpoint what caused that break. In contrast, the dummy variable approach to studying structural changes can pinpoint the exact cause of the break—intercept or slope or an interaction variable. This limit is also reflected in the fact that unless we know the exact break point, the Chow test cannot be performed. The dummy variable approach is more flexible as it can incorporate discontinuous periods of time and test across multiple break points.

Simply put, while the Chow Test is a straightforward way to test for a structural break at a known date, the Dummy Variable approach is more versatile, pinpointing the nature of the change, and useful even when the breakpoint's date is unknown.

5.1.11 Summary

Categorical variables, which lack a natural numerical scale, cannot be directly included in linear regression models designed to handle numerical data. Dummy variables, also known as binary, indicator, or categorical variables, provide a solution to this issue by encoding categorical information into a format suitable for regression analysis. By assigning values of 0 and 1 to represent the presence or absence of a particular category, dummy variables act as a bridge between qualitative data and quantitative models.

One of the primary applications of dummy variables is accounting for the impact of categorical factors on the outcome variable of interest. For instance, during periods of economic crises, the relationship between gross domestic consumption and national income may shift due to external forces. By creating a dummy variable that takes the value of 1 during crisis years and 0 during normal years, researchers can incorporate the effect of these episodes into their regression models.

Dummy variables can be employed to capture changes in either the intercept or the slope of the regression line. If economic crises are hypothesized to affect the level of autonomous consumption, the dummy variable can be used to model a shift in the intercept parameter. Alternatively, if the crises are expected to influence the marginal propensity to consume, the dummy variable can be included as an interaction term with the income variable, thereby modifying the slope parameter.

In addition to modeling economic events, dummy variables find applications in various other contexts. They can be used to control for seasonality in time series data, account for qualitative characteristics such as gender and profession in consumption or demand models, and test for interaction effects between categorical and continuous variables.

When dealing with categorical variables with more than two categories, researchers must exercise caution to avoid the "dummy variable trap," which occurs when dummy variables are created for all possible states, leading to exact multicollinearity. To circumvent this issue, it is recommended to

create $n - 1$ dummy variables for a categorical variable with n categories, where one category is designated as the reference level.

Dummy variables can be employed in conjunction with continuous quantitative variables to investigate more complex relationships. For instance, in a study examining the risk factors for cardiovascular events, a dummy variable could be created to represent high blood pressure, allowing researchers to assess its impact on the likelihood of experiencing such events while controlling for other factors like age and cholesterol levels.

An alternative approach to using dummy variables for detecting structural changes in regression relationships across different samples or time periods is the Chow test. While the Chow test provides an omnibus test for the presence of a structural break, it does not pinpoint the specific nature of the change. In contrast, the dummy variable approach offers greater versatility by allowing researchers to identify whether the structural change manifests as a shift in the intercept, a change in the slope, or the presence of an interaction effect.

5.1.12 Keywords

Dummy Variables: Binary (0/1) variables used to represent categorical data in regression models.

Categorical Variables: Variables that lack a natural numerical scale, such as gender, profession, or industry type.

Intercept Dummy: A dummy variable used to model a shift in the intercept of a regression line due to a categorical factor.

Slope Dummy: A dummy variable used as an interaction term to model a change in the slope coefficient due to a categorical factor.

Dummy Variable Trap: The situation of exact multicollinearity that arises when dummy variables are created for all categories of a variable.

Seasonal Dummies: Dummy variables used to control for seasonal patterns or trends in time series data.

Interaction Variable: The product of a dummy variable and a continuous variable, used to model interaction effects between categorical and quantitative variables.

Chow Test: A statistical test used to detect structural breaks or changes in regression relationships across different samples or time periods.

Confounding Variables: Variables that can influence the relationship between the dependent and independent variables, which need to be controlled for using dummy variables.

Treatment Effects Model: A regression model that uses dummy variables to represent both time (before and after treatment) and group membership (treatment or control group) to study the effects of a treatment or intervention.

5.1.13 Self-assessment Questions

1. Explain the concept of a dummy variable and its use in regression analysis. Provide an example from economics or finance where dummy variables would be useful.
2. Discuss the interpretation of the coefficients associated with dummy variables in a linear regression model. How does this interpretation differ from that of coefficients associated with continuous variables?
3. What are the potential problems that can arise when using dummy variables in a regression model? How can these problems be addressed?
4. Explain the concept of a base category when using dummy variables. How is the base category chosen, and what are the implications of this choice?
5. Discuss the use of interaction terms involving dummy variables in regression analysis. Provide an example where such an interaction term would be useful in understanding the relationship between variables.
6. Explain the difference between using dummy variables as part of the main effects and using them as part of an interaction effect in a regression model. Provide an example from economics or finance to illustrate your explanation.
7. Discuss the concept of dummy variable trap and how it can be avoided when including multiple dummy variables in a regression model.
8. In what situations would you recommend using dummy variables instead of continuous variables in a regression model? What are the potential advantages and disadvantages of this approach?

9. Explain how the interpretation of a dummy variable coefficient changes when the base category is changed. Provide an example to illustrate your explanation.
10. Discuss the use of nested dummy variables in regression analysis. When would this approach be useful, and what are the potential limitations?
11. Consider the following data on housing prices (in thousands of dollars) and various characteristics of houses in a city:

Price	Bedrooms	Bathrooms	Age	Garage
485	3	3	13	0
457	3	1	17	1
475	3	2	10	0
461	2	3	26	0
393	2	3	27	1
489	4	3	9	0
480	2	2	7	1
361	4	3	29	0
363	5	2	25	1
490	5	1	13	0
426	5	3	20	0
474	4	3	8	0
324	5	3	28	1
451	2	2	26	0
302	2	1	25	0
403	4	2	18	0
413	4	3	16	0
319	5	1	8	1
419	4	1	23	1
413	4	3	22	0

Price	Bedrooms	Bathrooms	Age	Garage
473	4	2	29	0
385	4	1	25	1
448	3	1	24	1
378	5	3	11	0
471	4	2	13	1
474	2	3	28	0
326	5	2	22	1
445	5	2	8	1
417	3	2	22	0
315	2	2	30	1

Estimate a multiple regression model with the housing price as the dependent variable and the other variables as independent variables, including a dummy variable for the presence of a garage. Test the significance of the garage dummy variable coefficient.

12. Using the same data from Question 11, estimate a regression model that includes interaction terms between the garage dummy variable and the number of bedrooms and bathrooms. Interpret the coefficients of the interaction terms.
13. Consider the following data on weekly earnings (in dollars) and various characteristics of individuals in a city:

Earnings	Education	Experience	Gender
862	18	19	Male
653	15	9	Male
999	14	17	Female
664	18	14	Male
770	12	8	Male
739	13	19	Female
874	16	11	Male

Earnings	Education	Experience	Gender
924	18	19	Female
997	15	20	Female
654	13	9	Male
683	17	18	Male
704	20	17	Male
663	15	19	Female
636	16	11	Male
654	14	12	Female
836	14	9	Female
880	12	5	Male
632	18	17	Female
683	14	20	Male
908	15	18	Male
838	20	20	Male
883	12	20	Female
864	15	15	Female
843	18	9	Female
935	15	20	Female
977	16	14	Male
687	17	4	Male
670	17	14	Female
688	12	17	Female
942	16	8	Female

Estimate a multiple regression model with earnings as the dependent variable and the other variables as independent variables, including a dummy variable for gender. Test the significance of the gender dummy variable coefficient.

14. Using the same data from Question 13, estimate a regression model that includes an interaction term between the gender dummy variable and the education level. Interpret the coefficient of the interaction term.
15. Consider the following data on stock returns (in percentage) and various characteristics of companies:

Company	Return	Size	Leverage	Industry
22	7	Large	High	Tech
7	16	Small	Low	Finance
12	13	Small	Low	Tech
20	7	Large	Low	Tech
29	17	Small	High	Finance
23	11	Large	Low	Finance
9	13	Large	High	Finance
13	12	Large	Low	Tech
3	15	Large	High	Finance
16	14	Small	High	Finance
17	18	Small	Low	Tech
25	20	Small	High	Tech
30	14	Large	Low	Manufacturing
26	18	Large	High	Finance
27	20	Small	High	Tech
11	8	Small	Low	Tech
14	6	Large	High	Finance
24	5	Small	High	Finance
4	5	Small	Low	Tech
6	10	Large	Low	Finance
2	8	Small	Low	Manufacturing

Company	Return	Size	Leverage	Industry
10	8	Large	Low	Finance
19	5	Large	High	Tech
8	15	Small	High	Tech
1	12	Large	High	Tech
18	5	Large	Low	Manufacturing
28	5	Small	High	Manufacturing
5	10	Large	High	Manufacturing
21	12	Large	High	Finance
15	7	Small	High	Tech

Estimate a multiple regression model with stock returns as the dependent variable and the other variables as independent variables, including dummy variables for company size, leverage, and industry. Test the joint significance of the industry dummy variable coefficients.

16. Consider the following data on monthly sales (in thousands of dollars) of a retail store and various factors:

Sales	Promotion	Season	Location
123	0	Spring	Suburban
199	1	Summer	Urban
179	1	Summer	Suburban
109	0	Winter	Suburban
193	0	Winter	Urban
100	0	Spring	Urban
133	0	Winter	Suburban
131	0	Summer	Urban
166	1	Winter	Urban
118	0	Fall	Suburban

Sales	Promotion	Season	Location
127	1	Winter	Suburban
135	0	Spring	Suburban
149	1	Winter	Suburban
138	0	Fall	Urban
111	0	Winter	Urban
163	0	Winter	Urban
181	1	Spring	Suburban
155	1	Summer	Suburban
161	1	Fall	Urban
121	0	Fall	Urban
109	0	Winter	Suburban
161	0	Fall	Suburban
135	0	Fall	Urban
178	1	Summer	Urban
108	0	Summer	Urban
191	1	Winter	Suburban
123	1	Winter	Urban
129	1	Summer	Urban
177	0	Fall	Suburban
117	0	Fall	Suburban

Estimate a multiple regression model with sales as the dependent variable and the other variables as independent variables, including dummy variables for promotion, season, and location. Test the joint significance of the season dummy variable coefficients.

17. Using the same data from Question 16, estimate a regression model that includes interaction terms between the promotion dummy variable and the season dummy variables. Interpret the coefficients of the interaction terms.

18. Consider the following data on hourly wages (in dollars) and various characteristics of employees in a firm:

Wage	Education	Experience	Gender	Department
32	22	11	Male	Finance
22	19	15	Male	Finance
34	20	5	Male	Sales
25	20	4	Male	Sales
20	17	11	Male	Marketing
24	21	15	Female	Marketing
29	14	14	Male	Marketing
22	15	15	Female	Marketing
24	18	7	Female	Marketing
32	22	14	Female	Sales
32	21	13	Male	Marketing
25	21	5	Female	Finance
20	16	2	Female	Finance
35	16	13	Male	Finance
30	20	12	Male	Marketing
21	17	13	Male	Marketing
31	16	5	Male	Sales
29	18	6	Female	Marketing
35	18	2	Female	Finance
32	18	9	Female	Sales
33	19	8	Female	Finance
35	15	4	Female	Sales
20	21	13	Male	Sales
30	18	9	Male	Sales

Wage	Education	Experience	Gender	Department
30	19	8	Male	Sales
35	15	4	Female	Sales
33	22	14	Female	Finance
22	20	13	Male	Finance
23	19	10	Female	Finance
31	15	9	Female	Marketing

Estimate a multiple regression model with wages as the dependent variable and the other variables as independent variables, including dummy variables for gender and department. Test the significance of the gender dummy variable coefficient.

19. Using the same data from Question 18, estimate a regression model that includes an interaction term between the gender dummy variable and the experience level. Interpret the coefficient of the interaction term.
20. Consider the following data on customer satisfaction scores (on a scale of 1 to 10) and various characteristics of restaurants:

Satisfaction	Price	Service	Cuisine
5	Low	Poor	Italian
10	High	Poor	Mexican
7	Low	Excellent	Italian
8	High	Poor	Mexican
5	Low	Good	Chinese
7	High	Excellent	Mexican
6	High	Good	Italian
10	Low	Good	Chinese
7	Low	Excellent	Chinese
6	High	Good	Italian
5	Low	Good	Mexican

Satisfaction	Price	Service	Cuisine
10	High	Excellent	French
8	Low	Poor	Italian
7	Low	Good	Italian
6	High	Good	Mexican
8	High	Excellent	Italian
10	High	Good	Mexican
9	High	Poor	Chinese
8	High	Poor	Italian
8	Low	Good	Italian
7	Low	Poor	French
10	Low	Poor	French
5	Low	Excellent	Italian
5	High	Poor	Mexican
9	High	Excellent	French
6	Low	Excellent	Chinese
9	Low	Poor	Italian
8	High	Excellent	French
9	High	Good	Mexican
6	Low	Excellent	Mexican

Estimate a multiple regression model with customer satisfaction as the dependent variable and the other variables as independent variables, including dummy variables for price, service, and cuisine. Test the joint significance of the cuisine dummy variable coefficients.

5.1.14 References

1. **Econometric Analysis** by William H. Greene (Pearson, 8th Edition, 2018) - A graduate-level textbook that provides in-depth treatment of dummy variable models, interaction effects, and various econometric techniques.
2. **Applied Econometrics** by Dimitrios Asteriou and Stephen G. Hall (Palgrave Macmillan, 4th Edition, 2022) - Covers the use of dummy variables in various economic and financial applications, with a focus on practical implementation and interpretation.
3. **Econometric Models and Economic Forecasts** by Robert S. Pindyck and Daniel L. Rubinfeld (McGraw-Hill Education, 9th Edition, 2017) - Discusses the use of dummy variables in forecasting models and time series analysis, as well as their applications in policy evaluation.
4. **Econometric Theory and Methods** by Russell Davidson and James G. MacKinnon (Oxford University Press, 2004) - A comprehensive reference on econometric theory, including advanced topics related to dummy variable models and structural change analysis.
5. **Econometric Analysis of Cross Section and Panel Data** by Jeffrey M. Wooldridge (MIT Press, 2010) - Focuses on the use of dummy variables in cross-sectional and panel data analysis, with applications in various fields of economics.

APPENDIX

Statistical Tables

z	*−0.00*	*−0.01*	*−0.02*	*−0.03*	*−0.04*	*−0.05*	*−0.06*	*−0.07*	*−0.08*	*−0.09*
−4.0	0.00003	0.00003	0.00003	0.00003	0.00003	0.00003	0.00002	0.00002	0.00002	0.00002
−3.9	0.00005	0.00005	0.00004	0.00004	0.00004	0.00004	0.00004	0.00004	0.00003	0.00003
−3.8	0.00007	0.00007	0.00007	0.00006	0.00006	0.00006	0.00006	0.00005	0.00005	0.00005
−3.7	0.00011	0.00010	0.00010	0.00010	0.00009	0.00009	0.00008	0.00008	0.00008	0.00008
−3.6	0.00016	0.00015	0.00015	0.00014	0.00014	0.00013	0.00013	0.00012	0.00012	0.00011
−3.5	0.00023	0.00022	0.00022	0.00021	0.00020	0.00019	0.00019	0.00018	0.00017	0.00017
−3.4	0.00034	0.00032	0.00031	0.00030	0.00029	0.00028	0.00027	0.00026	0.00025	0.00024
−3.3	0.00048	0.00047	0.00045	0.00043	0.00042	0.00040	0.00039	0.00038	0.00036	0.00035
−3.2	0.00069	0.00066	0.00064	0.00062	0.00060	0.00058	0.00056	0.00054	0.00052	0.00050
−3.1	0.00097	0.00094	0.00090	0.00087	0.00084	0.00082	0.00079	0.00076	0.00074	0.00071
−3.0	0.00135	0.00131	0.00126	0.00122	0.00118	0.00114	0.00111	0.00107	0.00104	0.00100
−2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
−2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
−2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
−2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
−2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
−2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
−2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
−2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
−2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
−2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
−1.9	0.02872	0.02807	0.02743	0.02680	0.02619	0.02559	0.02500	0.02442	0.02385	0.02330
−1.8	0.03593	0.03515	0.03438	0.03362	0.03288	0.03216	0.03144	0.03074	0.03005	0.02938
−1.7	0.04457	0.04363	0.04272	0.04182	0.04093	0.04006	0.03920	0.03836	0.03754	0.03673
−1.6	0.05480	0.05370	0.05262	0.05155	0.05050	0.04947	0.04846	0.04746	0.04648	0.04551

−1.5	0.06681	0.06552	0.06426	0.06301	0.06178	0.06057	0.05938	0.05821	0.05705	0.05592
−1.4	0.08076	0.07927	0.07780	0.07636	0.07493	0.07353	0.07215	0.07078	0.06944	0.06811
−1.3	0.09680	0.09510	0.09342	0.09176	0.09012	0.08851	0.08692	0.08534	0.08379	0.08226
−1.2	0.11507	0.11314	0.11123	0.10935	0.10749	0.10565	0.10383	0.10204	0.10027	0.09853
−1.1	0.13567	0.13350	0.13136	0.12924	0.12714	0.12507	0.12302	0.12100	0.11900	0.11702
−1.0	0.15866	0.15625	0.15386	0.15151	0.14917	0.14686	0.14457	0.14231	0.14007	0.13786
−0.9	0.18406	0.18141	0.17879	0.17619	0.17361	0.17106	0.16853	0.16602	0.16354	0.16109
−0.8	0.21186	0.20897	0.20611	0.20327	0.20045	0.19766	0.19489	0.19215	0.18943	0.18673
−0.7	0.24196	0.23885	0.23576	0.23270	0.22965	0.22663	0.22363	0.22065	0.21770	0.21476
−0.6	0.27425	0.27093	0.26763	0.26435	0.26109	0.25785	0.25463	0.25143	0.24825	0.24510
−0.5	0.30854	0.30503	0.30153	0.29806	0.29460	0.29116	0.28774	0.28434	0.28096	0.27760
−0.4	0.34458	0.34090	0.33724	0.33360	0.32997	0.32636	0.32276	0.31918	0.31561	0.31207
−0.3	0.38209	0.37828	0.37448	0.37070	0.36693	0.36317	0.35942	0.35569	0.35197	0.34827
−0.2	0.42074	0.41683	0.41294	0.40905	0.40517	0.40129	0.39743	0.39358	0.38974	0.38591
−0.1	0.46017	0.45620	0.45224	0.44828	0.44433	0.44038	0.43644	0.43251	0.42858	0.42465
−0.0	0.50000	0.49601	0.49202	0.48803	0.48405	0.48006	0.47608	0.47210	0.46812	0.46414

z	***0***	***0.01***	***0.02***	***0.03***	***0.04***	***0.05***	***0.06***	***0.07***	***0.08***	***0.09***
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56360	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361

2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99992	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998

One-sided	***75%***	***80%***	***85%***	***90%***	***95%***	***97.50%***	***99%***	***99.50%***	***99.75%***	***99.90%***	***99.95%***
Two-sided	***50%***	***60%***	***70%***	***80%***	***90%***	***95%***	***98%***	***99%***	***99.50%***	***99.80%***	***99.90%***
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767

24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	2.871	3.174	3.390
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291
One-sided	***75%***	***80%***	***85%***	***90%***	***95%***	***97.50%***	***99%***	***99.50%***	***99.75%***	***99.90%***	***99.95%***
Two-sided	***50%***	***60%***	***70%***	***80%***	***90%***	***95%***	***98%***	***99%***	***99.50%***	***99.80%***	***99.90%***

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315
21	8.034	10.283	26.171	29.615	32.671	35.479	36.343	38.932	41.401	44.522	46.797
22	8.643	10.982	27.301	30.813	33.924	36.781	37.659	40.289	42.796	45.962	48.268
23	9.26	11.689	28.429	32.007	35.172	38.076	38.968	41.638	44.181	47.391	49.728
24	9.886	12.401	29.553	33.196	36.415	39.364	40.27	42.98	45.559	48.812	51.179

25	10.52	13.12	30.675	34.382	37.652	40.646	41.566	44.314	46.928	50.223	52.62
26	11.16	13.844	31.795	35.563	38.885	41.923	42.856	45.642	48.29	51.627	54.052
27	11.808	14.573	32.912	36.741	40.113	43.195	44.14	46.963	49.645	53.023	55.476
28	12.461	15.308	34.027	37.916	41.337	44.461	45.419	48.278	50.993	54.411	56.892
29	13.121	16.047	35.139	39.087	42.557	45.722	46.693	49.588	52.336	55.792	58.301
30	13.787	16.791	36.25	40.256	43.773	46.979	47.962	50.892	53.672	57.167	59.703
31	14.458	17.539	37.359	41.422	44.985	48.232	49.226	52.191	55.003	58.536	61.098
32	15.134	18.291	38.466	42.585	46.194	49.48	50.487	53.486	56.328	59.899	62.487
33	15.815	19.047	39.572	43.745	47.4	50.725	51.743	54.776	57.648	61.256	63.87
34	16.501	19.806	40.676	44.903	48.602	51.966	52.995	56.061	58.964	62.608	65.247
35	17.192	20.569	41.778	46.059	49.802	53.203	54.244	57.342	60.275	63.955	66.619
36	17.887	21.336	42.879	47.212	50.998	54.437	55.489	58.619	61.581	65.296	67.985
37	18.586	22.106	43.978	48.363	52.192	55.668	56.73	59.893	62.883	66.633	69.346
38	19.289	22.878	45.076	49.513	53.384	56.896	57.969	61.162	64.181	67.966	70.703
39	19.996	23.654	46.173	50.66	54.572	58.12	59.204	62.428	65.476	69.294	72.055
40	20.707	24.433	47.269	51.805	55.758	59.342	60.436	63.691	66.766	70.618	73.402

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
41	21.421	25.215	48.363	52.949	56.942	60.561	61.665	64.95	68.053	71.938	74.745
42	22.138	25.999	49.456	54.09	58.124	61.777	62.892	66.206	69.336	73.254	76.084
43	22.859	26.785	50.548	55.23	59.304	62.99	64.116	67.459	70.616	74.566	77.419
44	23.584	27.575	51.639	56.369	60.481	64.201	65.337	68.71	71.893	75.874	78.75
45	24.311	28.366	52.729	57.505	61.656	65.41	66.555	69.957	73.166	77.179	80.077
46	25.041	29.16	53.818	58.641	62.83	66.617	67.771	71.201	74.437	78.481	81.4
47	25.775	29.956	54.906	59.774	64.001	67.821	68.985	72.443	75.704	79.78	82.72
48	26.511	30.755	55.993	60.907	65.171	69.023	70.197	73.683	76.969	81.075	84.037
49	27.249	31.555	57.079	62.038	66.339	70.222	71.406	74.919	78.231	82.367	85.351
50	27.991	32.357	58.164	63.167	67.505	71.42	72.613	76.154	79.49	83.657	86.661
51	28.735	33.162	59.248	64.295	68.669	72.616	73.818	77.386	80.747	84.943	87.968
52	29.481	33.968	60.332	65.422	69.832	73.81	75.021	78.616	82.001	86.227	89.272
53	30.23	34.776	61.414	66.548	70.993	75.002	76.223	79.843	83.253	87.507	90.573
54	30.981	35.586	62.496	67.673	72.153	76.192	77.422	81.069	84.502	88.786	91.872
55	31.735	36.398	63.577	68.796	73.311	77.38	78.619	82.292	85.749	90.061	93.168
56	32.49	37.212	64.658	69.919	74.468	78.567	79.815	83.513	86.994	91.335	94.461
57	33.248	38.027	65.737	71.04	75.624	79.752	81.009	84.733	88.236	92.605	95.751
58	34.008	38.844	66.816	72.16	76.778	80.936	82.201	85.95	89.477	93.874	97.039
59	34.77	39.662	67.894	73.279	77.931	82.117	83.391	87.166	90.715	95.14	98.324
60	35.534	40.482	68.972	74.397	79.082	83.298	84.58	88.379	91.952	96.404	99.607
61	36.301	41.303	70.049	75.514	80.232	84.476	85.767	89.591	93.186	97.665	100.888
62	37.068	42.126	71.125	76.63	81.381	85.654	86.953	90.802	94.419	98.925	102.166
63	37.838	42.95	72.201	77.745	82.529	86.83	88.137	92.01	95.649	100.182	103.442
64	38.61	43.776	73.276	78.86	83.675	88.004	89.32	93.217	96.878	101.437	104.716

65	39.383	44.603	74.351	79.973	84.821	89.177	90.501	94.422	98.105	102.691	105.988
66	40.158	45.431	75.424	81.085	85.965	90.349	91.681	95.626	99.33	103.942	107.258
67	40.935	46.261	76.498	82.197	87.108	91.519	92.86	96.828	100.554	105.192	108.526
68	41.713	47.092	77.571	83.308	88.25	92.689	94.037	98.028	101.776	106.44	109.791
69	42.494	47.924	78.643	84.418	89.391	93.856	95.213	99.228	102.996	107.685	111.055
70	43.275	48.758	79.715	85.527	90.531	95.023	96.388	100.425	104.215	108.929	112.317
71	44.058	49.592	80.786	86.635	91.67	96.189	97.561	101.621	105.432	110.172	113.577
72	44.843	50.428	81.857	87.743	92.808	97.353	98.733	102.816	106.648	111.412	114.835
73	45.629	51.265	82.927	88.85	93.945	98.516	99.904	104.01	107.862	112.651	116.092
74	46.417	52.103	83.997	89.956	95.081	99.678	101.074	105.202	109.074	113.889	117.346
75	47.206	52.942	85.066	91.061	96.217	100.839	102.243	106.393	110.286	115.125	118.599
76	47.997	53.782	86.135	92.166	97.351	101.999	103.41	107.583	111.495	116.359	119.85
77	48.788	54.623	87.203	93.27	98.484	103.158	104.576	108.771	112.704	117.591	121.1
78	49.582	55.466	88.271	94.374	99.617	104.316	105.742	109.958	113.911	118.823	122.348
79	50.376	56.309	89.338	95.476	100.749	105.473	106.906	111.144	115.117	120.052	123.594
80	51.172	57.153	90.405	96.578	101.879	106.629	108.069	112.329	116.321	121.28	124.839

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
81	51.969	57.998	91.472	97.68	103.01	107.783	109.232	113.512	117.524	122.507	126.083
82	52.767	58.845	92.538	98.78	104.139	108.937	110.393	114.695	118.726	123.733	127.324
83	53.567	59.692	93.604	99.88	105.267	110.09	111.553	115.876	119.927	124.957	128.565
84	54.368	60.54	94.669	100.98	106.395	111.242	112.712	117.057	121.126	126.179	129.804
85	55.17	61.389	95.734	102.079	107.522	112.393	113.871	118.236	122.325	127.401	131.041
86	55.973	62.239	96.799	103.177	108.648	113.544	115.028	119.414	123.522	128.621	132.277
87	56.777	63.089	97.863	104.275	109.773	114.693	116.184	120.591	124.718	129.84	133.512
88	57.582	63.941	98.927	105.372	110.898	115.841	117.34	121.767	125.913	131.057	134.745
89	58.389	64.793	99.991	106.469	112.022	116.989	118.495	122.942	127.106	132.273	135.978
90	59.196	65.647	101.054	107.565	113.145	118.136	119.648	124.116	128.299	133.489	137.208
91	60.005	66.501	102.117	108.661	114.268	119.282	120.801	125.289	129.491	134.702	138.438
92	60.815	67.356	103.179	109.756	115.39	120.427	121.954	126.462	130.681	135.915	139.666
93	61.625	68.211	104.241	110.85	116.511	121.571	123.105	127.633	131.871	137.127	140.893
94	62.437	69.068	105.303	111.944	117.632	122.715	124.255	128.803	133.059	138.337	142.119
95	63.25	69.925	106.364	113.038	118.752	123.858	125.405	129.973	134.247	139.546	143.344
96	64.063	70.783	107.425	114.131	119.871	125	126.554	131.141	135.433	140.755	144.567
97	64.878	71.642	108.486	115.223	120.99	126.141	127.702	132.309	136.619	141.962	145.789
98	65.694	72.501	109.547	116.315	122.108	127.282	128.849	133.476	137.803	143.168	147.01
99	66.51	73.361	110.607	117.407	123.225	128.422	129.996	134.642	138.987	144.373	148.23
100	67.328	74.222	111.667	118.498	124.342	129.561	131.142	135.807	140.169	145.577	149.449
101	68.146	75.083	112.726	119.589	125.458	130.7	132.287	136.971	141.351	146.78	150.667
102	68.965	75.946	113.786	120.679	126.574	131.838	133.431	138.134	142.532	147.982	151.884

103	69.785	76.809	114.845	121.769	127.689	132.975	134.575	139.297	143.712	149.183	153.099
104	70.606	77.672	115.903	122.858	128.804	134.111	135.718	140.459	144.891	150.383	154.314
105	71.428	78.536	116.962	123.947	129.918	135.247	136.86	141.62	146.07	151.582	155.528
106	72.251	79.401	118.02	125.035	131.031	136.382	138.002	142.78	147.247	152.78	156.74
107	73.075	80.267	119.078	126.123	132.144	137.517	139.143	143.94	148.424	153.977	157.952
108	73.899	81.133	120.135	127.211	133.257	138.651	140.283	145.099	149.599	155.173	159.162
109	74.724	82	121.192	128.298	134.369	139.784	141.423	146.257	150.774	156.369	160.372
110	75.55	82.867	122.25	129.385	135.48	140.917	142.562	147.414	151.948	157.563	161.581
111	76.377	83.735	123.306	130.472	136.591	142.049	143.7	148.571	153.122	158.757	162.788
112	77.204	84.604	124.363	131.558	137.701	143.18	144.838	149.727	154.294	159.95	163.995
113	78.033	85.473	125.419	132.643	138.811	144.311	145.975	150.882	155.466	161.141	165.201
114	78.862	86.342	126.475	133.729	139.921	145.441	147.111	152.037	156.637	162.332	166.406
115	79.692	87.213	127.531	134.813	141.03	146.571	148.247	153.191	157.808	163.523	167.61
116	80.522	88.084	128.587	135.898	142.138	147.7	149.383	154.344	158.977	164.712	168.813
117	81.353	88.955	129.642	136.982	143.246	148.829	150.517	155.496	160.146	165.9	170.016
118	82.185	89.827	130.697	138.066	144.354	149.957	151.652	156.648	161.314	167.088	171.217
119	83.018	90.7	131.752	139.149	145.461	151.084	152.785	157.8	162.481	168.275	172.418
120	83.852	91.573	132.806	140.233	146.567	152.211	153.918	158.95	163.648	169.461	173.617

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
121	84.686	92.446	133.861	141.315	147.674	153.338	155.051	160.1	164.814	170.647	174.816
122	85.52	93.32	134.915	142.398	148.779	154.464	156.183	161.25	165.98	171.831	176.014
123	86.356	94.195	135.969	143.48	149.885	155.589	157.314	162.398	167.144	173.015	177.212
124	87.192	95.07	137.022	144.562	150.989	156.714	158.445	163.546	168.308	174.198	178.408
125	88.029	95.946	138.076	145.643	152.094	157.839	159.575	164.694	169.471	175.38	179.604
126	88.866	96.822	139.129	146.724	153.198	158.962	160.705	165.841	170.634	176.562	180.799
127	89.704	97.698	140.182	147.805	154.302	160.086	161.834	166.987	171.796	177.743	181.993
128	90.543	98.576	141.235	148.885	155.405	161.209	162.963	168.133	172.957	178.923	183.186
129	91.382	99.453	142.288	149.965	156.508	162.331	164.091	169.278	174.118	180.103	184.379
130	92.222	100.331	143.34	151.045	157.61	163.453	165.219	170.423	175.278	181.282	185.571
131	93.063	101.21	144.392	152.125	158.712	164.575	166.346	171.567	176.438	182.46	186.762
132	93.904	102.089	145.444	153.204	159.814	165.696	167.473	172.711	177.597	183.637	187.953
133	94.746	102.968	146.496	154.283	160.915	166.816	168.6	173.854	178.755	184.814	189.142
134	95.588	103.848	147.548	155.361	162.016	167.936	169.725	174.996	179.913	185.99	190.331
135	96.431	104.729	148.599	156.44	163.116	169.056	170.851	176.138	181.07	187.165	191.52
136	97.275	105.609	149.651	157.518	164.216	170.175	171.976	177.28	182.226	188.34	192.707
137	98.119	106.491	150.702	158.595	165.316	171.294	173.1	178.421	183.382	189.514	193.894
138	98.964	107.372	151.753	159.673	166.415	172.412	174.224	179.561	184.538	190.688	195.08
139	99.809	108.254	152.803	160.75	167.514	173.53	175.348	180.701	185.693	191.861	196.266

140	100.655	109.137	153.854	161.827	168.613	174.648	176.471	181.84	186.847	193.033	197.451
141	101.501	110.02	154.904	162.904	169.711	175.765	177.594	182.979	188.001	194.205	198.635
142	102.348	110.903	155.954	163.98	170.809	176.882	178.716	184.118	189.154	195.376	199.819
143	103.196	111.787	157.004	165.056	171.907	177.998	179.838	185.256	190.306	196.546	201.002
144	104.044	112.671	158.054	166.132	173.004	179.114	180.959	186.393	191.458	197.716	202.184
145	104.892	113.556	159.104	167.207	174.101	180.229	182.08	187.53	192.61	198.885	203.366
146	105.741	114.441	160.153	168.283	175.198	181.344	183.2	188.666	193.761	200.054	204.547
147	106.591	115.326	161.202	169.358	176.294	182.459	184.321	189.802	194.912	201.222	205.727
148	107.441	116.212	162.251	170.432	177.39	183.573	185.44	190.938	196.062	202.39	206.907
149	108.291	117.098	163.3	171.507	178.485	184.687	186.56	192.073	197.211	203.557	208.086
150	109.142	117.985	164.349	172.581	179.581	185.8	187.678	193.208	198.36	204.723	209.265
151	109.994	118.871	165.398	173.655	180.676	186.914	188.797	194.342	199.509	205.889	210.443
152	110.846	119.759	166.446	174.729	181.77	188.026	189.915	195.476	200.657	207.054	211.62
153	111.698	120.646	167.495	175.803	182.865	189.139	191.033	196.609	201.804	208.219	212.797
154	112.551	121.534	168.543	176.876	183.959	190.251	192.15	197.742	202.951	209.383	213.973
155	113.405	122.423	169.591	177.949	185.052	191.362	193.267	198.874	204.098	210.547	215.149
156	114.259	123.312	170.639	179.022	186.146	192.474	194.384	200.006	205.244	211.71	216.324
157	115.113	124.201	171.686	180.094	187.239	193.584	195.5	201.138	206.39	212.873	217.499
158	115.968	125.09	172.734	181.167	188.332	194.695	196.616	202.269	207.535	214.035	218.673
159	116.823	125.98	173.781	182.239	189.424	195.805	197.731	203.4	208.68	215.197	219.846
160	117.679	126.87	174.828	183.311	190.516	196.915	198.846	204.53	209.824	216.358	221.019

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
161	118.536	127.761	175.875	184.382	191.608	198.025	199.961	205.66	210.968	217.518	222.191
162	119.392	128.651	176.922	185.454	192.7	199.134	201.076	206.79	212.111	218.678	223.363
163	120.249	129.543	177.969	186.525	193.791	200.243	202.19	207.919	213.254	219.838	224.535
164	121.107	130.434	179.016	187.596	194.883	201.351	203.303	209.047	214.396	220.997	225.705
165	121.965	131.326	180.062	188.667	195.973	202.459	204.417	210.176	215.539	222.156	226.876
166	122.823	132.218	181.109	189.737	197.064	203.567	205.53	211.304	216.68	223.314	228.045
167	123.682	133.111	182.155	190.808	198.154	204.675	206.642	212.431	217.821	224.472	229.215
168	124.541	134.003	183.201	191.878	199.244	205.782	207.755	213.558	218.962	225.629	230.383
169	125.401	134.897	184.247	192.948	200.334	206.889	208.867	214.685	220.102	226.786	231.552
170	126.261	135.79	185.293	194.017	201.423	207.995	209.978	215.812	221.242	227.942	232.719
171	127.122	136.684	186.338	195.087	202.513	209.102	211.09	216.938	222.382	229.098	233.887
172	127.983	137.578	187.384	196.156	203.602	210.208	212.201	218.063	223.521	230.253	235.053
173	128.844	138.472	188.429	197.225	204.69	211.313	213.311	219.189	224.66	231.408	236.22
174	129.706	139.367	189.475	198.294	205.779	212.419	214.422	220.314	225.798	232.563	237.385
175	130.568	140.262	190.52	199.363	206.867	213.524	215.532	221.438	226.936	233.717	238.551
176	131.43	141.157	191.565	200.432	207.955	214.628	216.641	222.563	228.074	234.87	239.716
177	132.293	142.053	192.61	201.5	209.042	215.733	217.751	223.687	229.211	236.023	240.88
178	133.157	142.949	193.654	202.568	210.13	216.837	218.86	224.81	230.347	237.176	242.044

179	134.02	143.845	194.699	203.636	211.217	217.941	219.969	225.933	231.484	238.328	243.207
180	134.884	144.741	195.743	204.704	212.304	219.044	221.077	227.056	232.62	239.48	244.37
181	135.749	145.638	196.788	205.771	213.391	220.148	222.185	228.179	233.755	240.632	245.533
182	136.614	146.535	197.832	206.839	214.477	221.251	223.293	229.301	234.891	241.783	246.695
183	137.479	147.432	198.876	207.906	215.563	222.353	224.401	230.423	236.026	242.933	247.857
184	138.344	148.33	199.92	208.973	216.649	223.456	225.508	231.544	237.16	244.084	249.018
185	139.21	149.228	200.964	210.04	217.735	224.558	226.615	232.665	238.294	245.234	250.179
186	140.077	150.126	202.008	211.106	218.82	225.66	227.722	233.786	239.428	246.383	251.339
187	140.943	151.024	203.052	212.173	219.906	226.761	228.828	234.907	240.561	247.532	252.499
188	141.81	151.923	204.095	213.239	220.991	227.863	229.935	236.027	241.694	248.681	253.659
189	142.678	152.822	205.139	214.305	222.076	228.964	231.04	237.147	242.827	249.829	254.818
190	143.545	153.721	206.182	215.371	223.16	230.064	232.146	238.266	243.959	250.977	255.976
191	144.413	154.621	207.225	216.437	224.245	231.165	233.251	239.386	245.091	252.124	257.135
192	145.282	155.521	208.268	217.502	225.329	232.265	234.356	240.505	246.223	253.271	258.292
193	146.15	156.421	209.311	218.568	226.413	233.365	235.461	241.623	247.354	254.418	259.45
194	147.02	157.321	210.354	219.633	227.496	234.465	236.566	242.742	248.485	255.564	260.607
195	147.889	158.221	211.397	220.698	228.58	235.564	237.67	243.86	249.616	256.71	261.763
196	148.759	159.122	212.439	221.763	229.663	236.664	238.774	244.977	250.746	257.855	262.92
197	149.629	160.023	213.482	222.828	230.746	237.763	239.877	246.095	251.876	259.001	264.075
198	150.499	160.925	214.524	223.892	231.829	238.861	240.981	247.212	253.006	260.145	265.231
199	151.37	161.826	215.567	224.957	232.912	239.96	242.084	248.329	254.135	261.29	266.386
200	152.241	162.728	216.609	226.021	233.994	241.058	243.187	249.445	255.264	262.434	267.541

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
201	153.112	163.63	217.651	227.085	235.077	242.156	244.29	250.561	256.393	263.578	268.695
202	153.984	164.532	218.693	228.149	236.159	243.254	245.392	251.677	257.521	264.721	269.849
203	154.856	165.435	219.735	229.213	237.24	244.351	246.494	252.793	258.649	265.864	271.002
204	155.728	166.338	220.777	230.276	238.322	245.448	247.596	253.908	259.777	267.007	272.155
205	156.601	167.241	221.818	231.34	239.403	246.545	248.698	255.023	260.904	268.149	273.308
206	157.474	168.144	222.86	232.403	240.485	247.642	249.799	256.138	262.031	269.291	274.46
207	158.347	169.047	223.901	233.466	241.566	248.739	250.9	257.253	263.158	270.432	275.612
208	159.221	169.951	224.943	234.529	242.647	249.835	252.001	258.367	264.285	271.574	276.764
209	160.095	170.855	225.984	235.592	243.727	250.931	253.102	259.481	265.411	272.715	277.915
210	160.969	171.759	227.025	236.655	244.808	252.027	254.202	260.595	266.537	273.855	279.066
211	161.843	172.664	228.066	237.717	245.888	253.122	255.302	261.708	267.662	274.995	280.217
212	162.718	173.568	229.107	238.78	246.968	254.218	256.402	262.821	268.788	276.135	281.367
213	163.593	174.473	230.148	239.842	248.048	255.313	257.502	263.934	269.912	277.275	282.517
214	164.469	175.378	231.189	240.904	249.128	256.408	258.601	265.047	271.037	278.414	283.666
215	165.344	176.283	232.23	241.966	250.207	257.503	259.701	266.159	272.162	279.553	284.815
216	166.22	177.189	233.27	243.028	251.286	258.597	260.8	267.271	273.286	280.692	285.964
217	167.096	178.095	234.311	244.09	252.365	259.691	261.898	268.383	274.409	281.83	287.112
218	167.973	179.001	235.351	245.151	253.444	260.785	262.997	269.495	275.533	282.968	288.261

219	168.85	179.907	236.391	246.213	254.523	261.879	264.095	270.606	276.656	284.106	289.408
220	169.727	180.813	237.432	247.274	255.602	262.973	265.193	271.717	277.779	285.243	290.556
221	170.604	181.72	238.472	248.335	256.68	264.066	266.291	272.828	278.902	286.38	291.703
222	171.482	182.627	239.512	249.396	257.758	265.159	267.389	273.939	280.024	287.517	292.85
223	172.36	183.534	240.552	250.457	258.837	266.252	268.486	275.049	281.146	288.653	293.996
224	173.238	184.441	241.592	251.517	259.914	267.345	269.584	276.159	282.268	289.789	295.142
225	174.116	185.348	242.631	252.578	260.992	268.438	270.681	277.269	283.39	290.925	296.288
226	174.995	186.256	243.671	253.638	262.07	269.53	271.777	278.379	284.511	292.061	297.433
227	175.874	187.164	244.711	254.699	263.147	270.622	272.874	279.488	285.632	293.196	298.579
228	176.753	188.072	245.75	255.759	264.224	271.714	273.97	280.597	286.753	294.331	299.723
229	177.633	188.98	246.79	256.819	265.301	272.806	275.066	281.706	287.874	295.465	300.868
230	178.512	189.889	247.829	257.879	266.378	273.898	276.162	282.814	288.994	296.6	302.012
231	179.392	190.797	248.868	258.939	267.455	274.989	277.258	283.923	290.114	297.734	303.156
232	180.273	191.706	249.908	259.998	268.531	276.08	278.354	285.031	291.234	298.867	304.299
233	181.153	192.615	250.947	261.058	269.608	277.171	279.449	286.139	292.353	300.001	305.443
234	182.034	193.524	251.986	262.117	270.684	278.262	280.544	287.247	293.472	301.134	306.586
235	182.915	194.434	253.025	263.176	271.76	279.352	281.639	288.354	294.591	302.267	307.728
236	183.796	195.343	254.063	264.235	272.836	280.443	282.734	289.461	295.71	303.4	308.871
237	184.678	196.253	255.102	265.294	273.911	281.533	283.828	290.568	296.828	304.532	310.013
238	185.56	197.163	256.141	266.353	274.987	282.623	284.922	291.675	297.947	305.664	311.154
239	186.442	198.073	257.179	267.412	276.062	283.713	286.016	292.782	299.065	306.796	312.296
240	187.324	198.984	258.218	268.471	277.138	284.802	287.11	293.888	300.182	307.927	313.437

	P										
DF	***0.995***	***0.975***	***0.2***	***0.1***	***0.05***	***0.025***	***0.02***	***0.01***	***0.005***	***0.002***	***0.001***
241	188.207	199.894	259.256	269.529	278.213	285.892	288.204	294.994	301.3	309.058	314.578
242	189.09	200.805	260.295	270.588	279.288	286.981	289.298	296.1	302.417	310.189	315.718
243	189.973	201.716	261.333	271.646	280.362	288.07	290.391	297.206	303.534	311.32	316.859
244	190.856	202.627	262.371	272.704	281.437	289.159	291.484	298.311	304.651	312.45	317.999
245	191.739	203.539	263.409	273.762	282.511	290.248	292.577	299.417	305.767	313.58	319.138
246	192.623	204.45	264.447	274.82	283.586	291.336	293.67	300.522	306.883	314.71	320.278
247	193.507	205.362	265.485	275.878	284.66	292.425	294.762	301.626	307.999	315.84	321.417
248	194.391	206.274	266.523	276.935	285.734	293.513	295.855	302.731	309.115	316.969	322.556
249	195.276	207.186	267.561	277.993	286.808	294.601	296.947	303.835	310.231	318.098	323.694
250	196.161	208.098	268.599	279.05	287.882	295.689	298.039	304.94	311.346	319.227	324.832
300	240.663	253.912	320.397	331.789	341.395	349.874	352.425	359.906	366.844	375.369	381.425
350	285.608	300.064	372.051	384.306	394.626	403.723	406.457	414.474	421.9	431.017	437.488
400	330.903	346.482	423.59	436.649	447.632	457.305	460.211	468.724	476.606	486.274	493.132
450	376.483	393.118	475.035	488.849	500.456	510.67	513.736	522.717	531.026	541.212	548.432
500	422.303	439.936	526.401	540.93	553.127	563.852	567.07	576.493	585.207	595.882	603.446
550	468.328	486.91	577.701	592.909	605.667	616.878	620.241	630.084	639.183	650.324	658.215
600	514.529	534.019	628.943	644.8	658.094	669.769	673.27	683.516	692.982	704.568	712.771

650	560.885	581.245	680.134	696.614	710.421	722.542	726.176	736.807	746.625	758.639	767.141
700	607.38	628.577	731.28	748.359	762.661	775.211	778.972	789.974	800.131	812.556	821.347
750	653.997	676.003	782.386	800.043	814.822	827.785	831.67	843.029	853.514	866.336	875.404
800	700.725	723.513	833.456	851.671	866.911	880.275	884.279	895.984	906.786	919.991	929.329
850	747.554	771.099	884.492	903.249	918.937	932.689	936.808	948.848	959.957	973.534	983.133
900	794.475	818.756	935.499	954.782	970.904	985.032	989.263	1001.63	1013.036	1026.974	1036.826
950	841.48	866.477	986.478	1006.272	1022.816	1037.311	1041.651	1054.334	1066.031	1080.32	1090.418
1000	888.564	914.257	1037.431	1057.724	1074.679	1089.531	1093.977	1106.969	1118.948	1133.579	1143.917

F Distribution critical values for P=0.10

Denom DF	***Numerator DF***											
	1	***2***	***3***	***4***	***5***	***7***	***10***	***15***	***20***	***30***	***60***	***120***	***500***
1	39.864	49.5	53.593	55.833	57.24	58.906	60.195	61.22	61.74	62.265	62.794	63.061	63.264
2	8.5264	8.9999	9.1618	9.2434	9.2926	9.3491	9.3915	9.4248	9.4413	9.458	9.4745	9.4829	9.4893
3	5.5384	5.4624	5.3907	5.3426	5.3092	5.2661	5.2304	5.2003	5.1845	5.1681	5.1513	5.1425	5.1358
4	4.5448	4.3245	4.1909	4.1073	4.0505	3.979	3.9198	3.8704	3.8443	3.8175	3.7896	3.7753	3.7643
5	4.0605	3.7798	3.6194	3.5202	3.453	3.3679	3.2974	3.2379	3.2067	3.174	3.1402	3.1228	3.1094
7	3.5895	3.2575	3.074	2.9605	2.8833	2.785	2.7025	2.6322	2.5947	2.5555	2.5142	2.4927	2.4761
10	3.285	2.9244	2.7277	2.6054	2.5216	2.4139	2.3226	2.2434	2.2007	2.1554	2.1071	2.0818	2.0618
15	3.0731	2.6951	2.4898	2.3615	2.2729	2.1582	2.0593	1.9722	1.9243	1.8727	1.8168	1.7867	1.7629
20	2.9746	2.5893	2.3801	2.249	2.1582	2.0397	1.9368	1.845	1.7939	1.7383	1.6768	1.6432	1.6163
30	2.8808	2.4887	2.2761	2.1423	2.0493	1.9269	1.8195	1.7222	1.6674	1.6064	1.5376	1.499	1.4669
60	2.7911	2.3932	2.1774	2.0409	1.9457	1.8194	1.707	1.6034	1.5435	1.4756	1.3953	1.3476	1.306
120	2.7478	2.3473	2.13	1.9924	1.8959	1.7675	1.6523	1.545	1.4821	1.4094	1.3203	1.2646	1.2123
500	2.7157	2.3132	2.0947	1.9561	1.8588	1.7288	1.6115	1.5009	1.4354	1.3583	1.26	1.1937	1.1215

F Distribution critical values for P=0.05

Denom DF	***Numerator DF***											
	1	***2***	***3***	***4***	***5***	***7***	***10***	***15***	***20***	***30***	***60***	***120***	***500***
1	161.45	199.5	215.71	224.58	230.16	236.77	241.88	245.95	248.01	250.1	252.2	253.25	254.06
2	18.513	19	19.164	19.247	19.296	19.353	19.396	19.429	19.446	19.462	19.479	19.487	19.494
3	10.128	9.5522	9.2766	9.1172	9.0135	8.8867	8.7855	8.7028	8.6602	8.6165	8.572	8.5493	8.532
4	7.7086	6.9443	6.5915	6.3882	6.256	6.0942	5.9644	5.8579	5.8026	5.7458	5.6877	5.658	5.6352
5	6.6078	5.7862	5.4095	5.1922	5.0504	4.8759	4.7351	4.6187	4.5582	4.4958	4.4314	4.3985	4.3731
7	5.5914	4.7375	4.3469	4.1202	3.9715	3.7871	3.6366	3.5108	3.4445	3.3758	3.3043	3.2675	3.2388
10	4.9645	4.1028	3.7082	3.478	3.3259	3.1354	2.9782	2.845	2.7741	2.6996	2.621	2.5801	2.5482
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7066	2.5437	2.4035	2.3275	2.2467	2.1601	2.1141	2.0776
20	4.3512	3.4928	3.0983	2.866	2.7109	2.514	2.3479	2.2032	2.1241	2.0391	1.9463	1.8962	1.8563
30	4.1709	3.3159	2.9223	2.6896	2.5336	2.3343	2.1646	2.0149	1.9317	1.8408	1.7396	1.6835	1.6376
60	4.0012	3.1505	2.7581	2.5252	2.3683	2.1666	1.9927	1.8365	1.748	1.6492	1.5343	1.4672	1.4093
120	3.9201	3.0718	2.6802	2.4473	2.2898	2.0868	1.9104	1.7505	1.6587	1.5544	1.4289	1.3519	1.2804
500	3.8601	3.0137	2.6227	2.3898	2.232	2.0278	1.8496	1.6864	1.5917	1.482	1.3455	1.2552	1.1586

F Distribution critical values for P=0.02													
Denom DF	Numerator DF												
	1	2	3	4	5	7	10	15	20	30	60	120	500
1	1012.5	1249.5	1350.5	1405.8	1440.6	1481.8	1513.7	1539.1	1551.9	1564.9	1578	1584.6	1589.6
2	48.505	49	49.166	49.249	49.299	49.356	49.398	49.432	49.448	49.465	49.482	49.49	49.496
3	20.618	18.858	18.11	17.694	17.429	17.11	16.86	16.657	16.553	16.448	16.34	16.286	16.244
4	14.04	12.142	11.344	10.899	10.616	10.274	10.003	9.7828	9.6696	9.554	9.4359	9.376	9.33
5	11.323	9.4544	8.6702	8.233	7.953	7.6137	7.3438	7.1234	7.0094	6.8926	6.7728	6.7119	6.6649
7	8.9877	7.2026	6.4539	6.0347	5.7647	5.4354	5.1711	4.9531	4.8392	4.722	4.6007	4.5384	4.4902
10	7.6384	5.9336	5.2182	4.8157	4.5549	4.2347	3.975	3.758	3.6437	3.5245	3.3999	3.3354	3.285
15	6.773	5.1355	4.4475	4.0584	3.8052	3.4917	3.2345	3.0168	2.9003	2.7775	2.6468	2.578	2.5237
20	6.3907	4.7875	4.1134	3.7312	3.4817	3.1713	2.9149	2.6955	2.5771	2.4509	2.3148	2.2421	2.1841
30	6.0382	4.4695	3.8093	3.4339	3.1877	2.8803	2.6239	2.402	2.2805	2.1493	2.0047	1.9255	1.8611
60	5.7127	4.1785	3.5319	3.1633	2.9207	2.6157	2.3586	2.1326	2.0067	1.8676	1.7085	1.6169	1.5383
120	5.5594	4.0423	3.4026	3.0372	2.7963	2.4923	2.2347	2.0059	1.8769	1.7322	1.5613	1.4577	1.3629
500	5.4467	3.9428	3.3083	2.9453	2.7057	2.4024	2.1441	1.9128	1.7809	1.6307	1.4468	1.3273	1.2019

F Distribution critical values for P=0.01													
Denom DF	Numerator DF												
	1	2	3	4	5	7	10	15	20	30	60	120	500
1	4052.2	4999.5	5403.4	5624.6	5763.6	5928.4	6055.8	6157.3	6208.7	6260.6	6313	6339.4	6359.5
2	98.503	99	99.166	99.249	99.299	99.356	99.399	99.433	99.449	99.466	99.482	99.491	99.497
3	34.116	30.817	29.457	28.71	28.237	27.672	27.229	26.872	26.69	26.504	26.316	26.221	26.148
4	21.198	18	16.694	15.977	15.522	14.976	14.546	14.198	14.02	13.838	13.652	13.558	13.486
5	16.258	13.274	12.06	11.392	10.967	10.455	10.051	9.7222	9.5526	9.3793	9.202	9.1118	9.0424
7	12.246	9.5467	8.4513	7.8466	7.4605	6.9929	6.6201	6.3143	6.1554	5.992	5.8236	5.7373	5.6707
10	10.044	7.5594	6.5523	5.9944	5.6363	5.2001	4.8492	4.5582	4.4055	4.2469	4.0818	3.9964	3.9303
15	8.6831	6.3588	5.4169	4.8932	4.5557	4.1416	3.8049	3.5223	3.3719	3.2141	3.0471	2.9594	2.8906
20	8.096	5.8489	4.9382	4.4306	4.1027	3.6987	3.3682	3.088	2.9377	2.7785	2.6078	2.5167	2.4446
30	7.5624	5.3903	4.5098	4.0179	3.699	3.3046	2.9791	2.7002	2.5486	2.3859	2.2078	2.1108	2.0321
60	7.0771	4.9774	4.1259	3.6491	3.3388	2.953	2.6318	2.3522	2.1978	2.0284	1.8362	1.7264	1.6328
120	6.8509	4.7865	3.949	3.4795	3.1736	2.7918	2.472	2.1914	2.0345	1.86	1.6557	1.533	1.4215
500	6.6858	4.6479	3.821	3.3569	3.0539	2.6751	2.3564	2.0746	1.9152	1.7353	1.5175	1.3774	1.2317

F Distribution critical values for P=0.005													
Denom DF	Numerator DF												
	1	2	3	4	5	7	10	15	20	30	60	120	500
1	16211	19999	21615	22500	23056	23715	24224	24630	24836	25044	25253	25359	25439
2	198.5	199	199.17	199.25	199.3	199.36	199.4	199.43	199.45	199.47	199.48	199.49	199.5
3	55.552	49.799	47.467	46.195	45.392	44.434	43.686	43.085	42.777	42.466	42.149	41.989	41.867
4	31.333	26.284	24.259	23.155	22.456	21.622	20.967	20.438	20.167	19.891	19.611	19.468	19.359

5	22.785	18.314	16.53	15.556	14.94	14.2	13.618	13.146	12.903	12.656	12.402	12.274	12.175
7	16.235	12.404	10.882	10.05	9.5221	8.8853	8.3803	7.9677	7.7539	7.5345	7.3088	7.1933	7.1044
10	12.826	9.427	8.0807	7.3428	6.8723	6.3025	5.8467	5.4706	5.274	5.0705	4.8592	4.7501	4.6656
15	10.798	7.7007	6.476	5.8029	5.3721	4.8473	4.4235	4.0697	3.8826	3.6868	3.4802	3.3722	3.2874
20	9.9439	6.9865	5.8176	5.1744	4.7616	4.2569	3.847	3.502	3.3178	3.1234	2.9159	2.8058	2.7186
30	9.1796	6.3547	5.2387	4.6233	4.2275	3.7416	3.3439	3.0058	2.8231	2.6277	2.4151	2.2998	2.2066
60	8.4946	5.795	4.729	4.1399	3.7599	3.2911	2.9042	2.5705	2.3872	2.1874	1.9621	1.8341	1.7256
120	8.1789	5.5393	4.4972	3.9207	3.5482	3.0874	2.7052	2.3728	2.1882	1.984	1.7468	1.6055	1.4778
500	7.9498	5.3548	4.3304	3.7632	3.3963	2.9414	2.5625	2.2303	2.0441	1.8352	1.5844	1.4245	1.2595

F Distribution critical values for P=0.001

Denom DF	***Numerator DF***												
	1	***2***	***3***	***4***	***5***	***7***	***10***	***15***	***20***	***30***	***60***	***120***	***500***
1	405284	499999	540379	562500	576405	592873	605621	615764	620908	626099	631337	633972	635983
2	998.5	999	999.17	999.25	999.3	999.36	999.4	999.43	999.45	999.47	999.48	999.49	999.5
3	167.03	148.5	141.11	137.1	134.58	131.58	129.25	127.37	126.42	125.45	124.47	123.97	123.59
4	74.137	61.245	56.177	53.436	51.712	49.658	48.053	46.761	46.1	45.429	44.746	44.4	44.135
5	47.181	37.122	33.202	31.085	29.752	28.163	26.917	25.911	25.395	24.869	24.333	24.061	23.852
7	29.245	21.689	18.772	17.198	16.206	15.019	14.083	13.324	12.932	12.53	12.119	11.909	11.747
10	21.04	14.905	12.553	11.283	10.481	9.5174	8.7539	8.1288	7.8038	7.4688	7.1224	6.9443	6.8065
15	16.587	11.339	9.3352	8.2526	7.5673	6.7408	6.0808	5.5351	5.2484	4.9502	4.6378	4.4749	4.3478
20	14.819	9.9526	8.0984	7.096	6.4606	5.692	5.0753	4.5618	4.29	4.0051	3.703	3.5439	3.4184
30	13.293	8.7734	7.0544	6.1245	5.5339	4.8173	4.2389	3.7528	3.4928	3.2171	2.9197	2.7595	2.631
60	11.973	7.7678	6.1712	5.3067	4.7565	4.0864	3.5415	3.0781	2.8265	2.5549	2.2522	2.0821	1.939
120	11.38	7.3212	5.7814	4.9471	4.4157	3.7669	3.2372	2.7833	2.5345	2.2621	1.9502	1.7668	1.6027
500	10.957	7.0041	5.5056	4.6935	4.1757	3.5424	3.0234	2.5759	2.3282	2.0538	1.7292	1.526	1.3191

DW (α=5%)										
n	***k=1***		***k=2***		***k=3***		***k=4***		***k=5***	
	dL	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***	***dL***	***dU***
6	0.6102	1.4002								
7	0.6996	1.3564	0.4672	1.8964						
8	0.7629	1.3324	0.5591	1.7771	0.3674	2.2866				
9	0.8243	1.3199	0.6291	1.6993	0.4548	2.1282	0.2957	2.5881		
10	0.8791	1.3197	0.6972	1.6413	0.5253	2.0163	0.3760	2.4137	0.2427	2.8217
11	0.9273	1.3241	0.7580	1.6044	0.5948	1.9280	0.4441	2.2833	0.3155	2.6446
12	0.9708	1.3314	0.8122	1.5794	0.6577	1.8640	0.5120	2.1766	0.3796	2.5061
13	1.0097	1.3404	0.8612	1.5621	0.7147	1.8159	0.5745	2.0943	0.4445	2.3897
14	1.0450	1.3503	0.9054	1.5507	0.7667	1.7788	0.6321	2.0296	0.5052	2.2959
15	1.0770	1.3605	0.9455	1.5432	0.8140	1.7501	0.6852	1.9774	0.5620	2.2198
16	1.1062	1.3709	0.9820	1.5386	0.8572	1.7277	0.7340	1.9351	0.6150	2.1567
17	1.1330	1.3812	1.0154	1.5361	0.8968	1.7101	0.7790	1.9005	0.6641	2.1041

18	1.1576	1.3913	1.0461	1.5353	0.9331	1.6961	0.8204	1.8719	0.7098	2.0600
19	1.1804	1.4012	1.0743	1.5355	0.9666	1.6851	0.8588	1.8482	0.7523	2.0226
20	1.2015	1.4107	1.1004	1.5367	0.9976	1.6763	0.8943	1.8283	0.7918	1.9908
21	1.2212	1.4200	1.1246	1.5385	1.0262	1.6694	0.9272	1.8116	0.8286	1.9635
22	1.2395	1.4289	1.1471	1.5408	1.0529	1.6640	0.9578	1.7974	0.8629	1.9400
23	1.2567	1.4375	1.1682	1.5435	1.0778	1.6597	0.9864	1.7855	0.8949	1.9196
24	1.2728	1.4458	1.1878	1.5464	1.1010	1.6565	1.0131	1.7753	0.9249	1.9018
25	1.2879	1.4537	1.2063	1.5495	1.1228	1.6540	1.0381	1.7666	0.9530	1.8863
26	1.3022	1.4614	1.2236	1.5528	1.1432	1.6523	1.0616	1.7591	0.9794	1.8727
27	1.3157	1.4688	1.2399	1.5562	1.1624	1.6510	1.0836	1.7527	1.0042	1.8608
28	1.3284	1.4759	1.2553	1.5596	1.1805	1.6503	1.1044	1.7473	1.0276	1.8502
29	1.3405	1.4828	1.2699	1.5631	1.1976	1.6499	1.1241	1.7426	1.0497	1.8409
30	1.3520	1.4894	1.2837	1.5666	1.2138	1.6498	1.1426	1.7386	1.0706	1.8326
31	1.3630	1.4957	1.2969	1.5701	1.2292	1.6500	1.1602	1.7352	1.0904	1.8252
32	1.3734	1.5019	1.3093	1.5736	1.2437	1.6505	1.1769	1.7323	1.1092	1.8187
33	1.3834	1.5078	1.3212	1.5770	1.2576	1.6511	1.1927	1.7298	1.1270	1.8128
34	1.3929	1.5136	1.3325	1.5805	1.2707	1.6519	1.2078	1.7277	1.1439	1.8076
35	1.4019	1.5191	1.3433	1.5838	1.2833	1.6528	1.2221	1.7259	1.1601	1.8029
36	1.4107	1.5245	1.3537	1.5872	1.2953	1.6539	1.2358	1.7245	1.1755	1.7987
37	1.4190	1.5297	1.3635	1.5904	1.3068	1.6550	1.2489	1.7233	1.1901	1.7950
38	1.4270	1.5348	1.3730	1.5937	1.3177	1.6563	1.2614	1.7223	1.2042	1.7916
39	1.4347	1.5396	1.3821	1.5969	1.3283	1.6575	1.2734	1.7215	1.2176	1.7886
40	1.4421	1.5444	1.3908	1.6000	1.3384	1.6589	1.2848	1.7209	1.2305	1.7859
41	1.4493	1.5490	1.3992	1.6031	1.3480	1.6603	1.2958	1.7205	1.2428	1.7835
42	1.4562	1.5534	1.4073	1.6061	1.3573	1.6617	1.3064	1.7202	1.2546	1.7814
43	1.4628	1.5577	1.4151	1.6091	1.3663	1.6632	1.3166	1.7200	1.2660	1.7794
44	1.4692	1.5619	1.4226	1.6120	1.3749	1.6647	1.3263	1.7200	1.2769	1.7777
45	1.4754	1.5660	1.4298	1.6148	1.3832	1.6662	1.3357	1.7200	1.2874	1.7762
46	1.4814	1.5700	1.4368	1.6176	1.3912	1.6677	1.3448	1.7201	1.2976	1.7748
47	1.4872	1.5739	1.4435	1.6204	1.3989	1.6692	1.3535	1.7203	1.3073	1.7736
48	1.4928	1.5776	1.4500	1.6231	1.4064	1.6708	1.3619	1.7206	1.3167	1.7725
49	1.4982	1.5813	1.4564	1.6257	1.4136	1.6723	1.3701	1.7210	1.3258	1.7716
50	1.5035	1.5849	1.4625	1.6283	1.4206	1.6739	1.3779	1.7214	1.3346	1.7708
51	1.5086	1.5884	1.4684	1.6309	1.4273	1.6754	1.3855	1.7218	1.3431	1.7701
52	1.5135	1.5917	1.4741	1.6334	1.4339	1.6769	1.3929	1.7223	1.3512	1.7694
53	1.5183	1.5951	1.4797	1.6359	1.4402	1.6785	1.4000	1.7228	1.3592	1.7689
54	1.5230	1.5983	1.4851	1.6383	1.4464	1.6800	1.4069	1.7234	1.3669	1.7684
55	1.5276	1.6014	1.4903	1.6406	1.4523	1.6815	1.4136	1.7240	1.3743	1.7681
56	1.5320	1.6045	1.4954	1.6430	1.4581	1.6830	1.4201	1.7246	1.3815	1.7678
57	1.5363	1.6075	1.5004	1.6452	1.4637	1.6845	1.4264	1.7253	1.3885	1.7675
58	1.5405	1.6105	1.5052	1.6475	1.4692	1.6860	1.4325	1.7259	1.3953	1.7673

59	1.5446	1.6134	1.5099	1.6497	1.4745	1.6875	1.4385	1.7266	1.4019	1.7672
60	1.5485	1.6162	1.5144	1.6518	1.4797	1.6889	1.4443	1.7274	1.4083	1.7671
61	1.5524	1.6189	1.5189	1.6540	1.4847	1.6904	1.4499	1.7281	1.4146	1.7671
62	1.5562	1.6216	1.5232	1.6561	1.4896	1.6918	1.4554	1.7288	1.4206	1.7671
63	1.5599	1.6243	1.5274	1.6581	1.4943	1.6932	1.4607	1.7296	1.4265	1.7671
64	1.5635	1.6268	1.5315	1.6601	1.4990	1.6946	1.4659	1.7303	1.4322	1.7672
65	1.5670	1.6294	1.5355	1.6621	1.5035	1.6960	1.4709	1.7311	1.4378	1.7673
66	1.5704	1.6318	1.5395	1.6640	1.5079	1.6974	1.4758	1.7319	1.4433	1.7675
67	1.5738	1.6343	1.5433	1.6660	1.5122	1.6988	1.4806	1.7327	1.4486	1.7676
68	1.5771	1.6367	1.5470	1.6678	1.5164	1.7001	1.4853	1.7335	1.4537	1.7678
69	1.5803	1.6390	1.5507	1.6697	1.5205	1.7015	1.4899	1.7343	1.4588	1.7680
70	1.5834	1.6413	1.5542	1.6715	1.5245	1.7028	1.4943	1.7351	1.4637	1.7683
71	1.5865	1.6435	1.5577	1.6733	1.5284	1.7041	1.4987	1.7358	1.4685	1.7685
72	1.5895	1.6457	1.5611	1.6751	1.5323	1.7054	1.5029	1.7366	1.4732	1.7688
73	1.5924	1.6479	1.5645	1.6768	1.5360	1.7067	1.5071	1.7375	1.4778	1.7691
74	1.5953	1.6500	1.5677	1.6785	1.5397	1.7079	1.5112	1.7383	1.4822	1.7694
75	1.5981	1.6521	1.5709	1.6802	1.5432	1.7092	1.5151	1.7390	1.4866	1.7698
76	1.6009	1.6541	1.5740	1.6819	1.5467	1.7104	1.5190	1.7399	1.4909	1.7701
77	1.6036	1.6561	1.5771	1.6835	1.5502	1.7117	1.5228	1.7407	1.4950	1.7704
78	1.6063	1.6581	1.5801	1.6851	1.5535	1.7129	1.5265	1.7415	1.4991	1.7708
79	1.6089	1.6601	1.5830	1.6867	1.5568	1.7141	1.5302	1.7423	1.5031	1.7712
80	1.6114	1.6620	1.5859	1.6882	1.5600	1.7153	1.5337	1.7430	1.5070	1.7716
81	1.6139	1.6639	1.5888	1.6898	1.5632	1.7164	1.5372	1.7438	1.5109	1.7720
82	1.6164	1.6657	1.5915	1.6913	1.5663	1.7176	1.5406	1.7446	1.5146	1.7724
83	1.6188	1.6675	1.5942	1.6928	1.5693	1.7187	1.5440	1.7454	1.5183	1.7728
84	1.6212	1.6693	1.5969	1.6942	1.5723	1.7199	1.5472	1.7462	1.5219	1.7732
85	1.6235	1.6711	1.5995	1.6957	1.5752	1.7210	1.5505	1.7470	1.5254	1.7736
86	1.6258	1.6728	1.6021	1.6971	1.5780	1.7221	1.5536	1.7478	1.5289	1.7740
87	1.6280	1.6745	1.6046	1.6985	1.5808	1.7232	1.5567	1.7485	1.5322	1.7745
88	1.6302	1.6762	1.6071	1.6999	1.5836	1.7243	1.5597	1.7493	1.5356	1.7749
89	1.6324	1.6778	1.6095	1.7013	1.5863	1.7254	1.5627	1.7501	1.5388	1.7754
90	1.6345	1.6794	1.6119	1.7026	1.5889	1.7264	1.5656	1.7508	1.5420	1.7758
91	1.6366	1.6810	1.6143	1.7040	1.5915	1.7275	1.5685	1.7516	1.5452	1.7763
92	1.6387	1.6826	1.6166	1.7053	1.5941	1.7285	1.5713	1.7523	1.5482	1.7767
93	1.6407	1.6841	1.6188	1.7066	1.5966	1.7295	1.5741	1.7531	1.5513	1.7772
94	1.6427	1.6857	1.6211	1.7078	1.5991	1.7306	1.5768	1.7538	1.5542	1.7776
95	1.6447	1.6872	1.6233	1.7091	1.6015	1.7316	1.5795	1.7546	1.5572	1.7781
96	1.6466	1.6887	1.6254	1.7103	1.6039	1.7326	1.5821	1.7553	1.5600	1.7785
97	1.6485	1.6901	1.6275	1.7116	1.6063	1.7335	1.5847	1.7560	1.5628	1.7790
98	1.6504	1.6916	1.6296	1.7128	1.6086	1.7345	1.5872	1.7567	1.5656	1.7795
99	1.6522	1.6930	1.6317	1.7140	1.6108	1.7355	1.5897	1.7575	1.5683	1.7799

100	1.6540	1.6944	1.6337	1.7152	1.6131	1.7364	1.5922	1.7582	1.5710	1.7804
101	1.6558	1.6958	1.6357	1.7163	1.6153	1.7374	1.5946	1.7589	1.5736	1.7809
102	1.6576	1.6971	1.6376	1.7175	1.6174	1.7383	1.5969	1.7596	1.5762	1.7813
103	1.6593	1.6985	1.6396	1.7186	1.6196	1.7392	1.5993	1.7603	1.5788	1.7818
104	1.6610	1.6998	1.6415	1.7198	1.6217	1.7402	1.6016	1.7610	1.5813	1.7823
105	1.6627	1.7011	1.6433	1.7209	1.6237	1.7411	1.6038	1.7617	1.5837	1.7827
106	1.6644	1.7024	1.6452	1.7220	1.6258	1.7420	1.6061	1.7624	1.5861	1.7832
107	1.6660	1.7037	1.6470	1.7231	1.6277	1.7428	1.6083	1.7631	1.5885	1.7837
108	1.6676	1.7050	1.6488	1.7241	1.6297	1.7437	1.6104	1.7637	1.5909	1.7841
109	1.6692	1.7062	1.6505	1.7252	1.6317	1.7446	1.6125	1.7644	1.5932	1.7846
110	1.6708	1.7074	1.6523	1.7262	1.6336	1.7455	1.6146	1.7651	1.5955	1.7851
111	1.6723	1.7086	1.6540	1.7273	1.6355	1.7463	1.6167	1.7657	1.5977	1.7855
112	1.6738	1.7098	1.6557	1.7283	1.6373	1.7472	1.6187	1.7664	1.5999	1.7860
113	1.6753	1.7110	1.6574	1.7293	1.6391	1.7480	1.6207	1.7670	1.6021	1.7864
114	1.6768	1.7122	1.6590	1.7303	1.6410	1.7488	1.6227	1.7677	1.6042	1.7869
115	1.6783	1.7133	1.6606	1.7313	1.6427	1.7496	1.6246	1.7683	1.6063	1.7874
116	1.6797	1.7145	1.6622	1.7323	1.6445	1.7504	1.6265	1.7690	1.6084	1.7878
117	1.6812	1.7156	1.6638	1.7332	1.6462	1.7512	1.6284	1.7696	1.6105	1.7883
118	1.6826	1.7167	1.6653	1.7342	1.6479	1.7520	1.6303	1.7702	1.6125	1.7887
119	1.6839	1.7178	1.6669	1.7352	1.6496	1.7528	1.6321	1.7709	1.6145	1.7892
120	1.6853	1.7189	1.6684	1.7361	1.6513	1.7536	1.6339	1.7715	1.6164	1.7896
121	1.6867	1.7200	1.6699	1.7370	1.6529	1.7544	1.6357	1.7721	1.6184	1.7901
122	1.6880	1.7210	1.6714	1.7379	1.6545	1.7552	1.6375	1.7727	1.6203	1.7905
123	1.6893	1.7221	1.6728	1.7388	1.6561	1.7559	1.6392	1.7733	1.6222	1.7910
124	1.6906	1.7231	1.6743	1.7397	1.6577	1.7567	1.6409	1.7739	1.6240	1.7914
125	1.6919	1.7241	1.6757	1.7406	1.6592	1.7574	1.6426	1.7745	1.6258	1.7919
126	1.6932	1.7252	1.6771	1.7415	1.6608	1.7582	1.6443	1.7751	1.6276	1.7923
127	1.6944	1.7261	1.6785	1.7424	1.6623	1.7589	1.6460	1.7757	1.6294	1.7928
128	1.6957	1.7271	1.6798	1.7432	1.6638	1.7596	1.6476	1.7763	1.6312	1.7932
129	1.6969	1.7281	1.6812	1.7441	1.6653	1.7603	1.6492	1.7769	1.6329	1.7937
130	1.6981	1.7291	1.6825	1.7449	1.6667	1.7610	1.6508	1.7774	1.6346	1.7941
131	1.6993	1.7301	1.6838	1.7458	1.6682	1.7617	1.6523	1.7780	1.6363	1.7945
132	1.7005	1.7310	1.6851	1.7466	1.6696	1.7624	1.6539	1.7786	1.6380	1.7950
133	1.7017	1.7319	1.6864	1.7474	1.6710	1.7631	1.6554	1.7791	1.6397	1.7954
134	1.7028	1.7329	1.6877	1.7482	1.6724	1.7638	1.6569	1.7797	1.6413	1.7958
135	1.7040	1.7338	1.6889	1.7490	1.6738	1.7645	1.6584	1.7802	1.6429	1.7962
136	1.7051	1.7347	1.6902	1.7498	1.6751	1.7652	1.6599	1.7808	1.6445	1.7967
137	1.7062	1.7356	1.6914	1.7506	1.6765	1.7659	1.6613	1.7813	1.6461	1.7971
138	1.7073	1.7365	1.6926	1.7514	1.6778	1.7665	1.6628	1.7819	1.6476	1.7975
139	1.7084	1.7374	1.6938	1.7521	1.6791	1.7672	1.6642	1.7824	1.6491	1.7979
140	1.7095	1.7382	1.6950	1.7529	1.6804	1.7678	1.6656	1.7830	1.6507	1.7984

141	1.7106	1.7391	1.6962	1.7537	1.6817	1.7685	1.6670	1.7835	1.6522	1.7988
142	1.7116	1.7400	1.6974	1.7544	1.6829	1.7691	1.6684	1.7840	1.6536	1.7992
143	1.7127	1.7408	1.6985	1.7552	1.6842	1.7697	1.6697	1.7846	1.6551	1.7996
144	1.7137	1.7417	1.6996	1.7559	1.6854	1.7704	1.6710	1.7851	1.6565	1.8000
145	1.7147	1.7425	1.7008	1.7566	1.6866	1.7710	1.6724	1.7856	1.6580	1.8004
146	1.7157	1.7433	1.7019	1.7574	1.6878	1.7716	1.6737	1.7861	1.6594	1.8008
147	1.7167	1.7441	1.7030	1.7581	1.6890	1.7722	1.6750	1.7866	1.6608	1.8012
148	1.7177	1.7449	1.7041	1.7588	1.6902	1.7729	1.6762	1.7871	1.6622	1.8016
149	1.7187	1.7457	1.7051	1.7595	1.6914	1.7735	1.6775	1.7876	1.6635	1.8020
150	1.7197	1.7465	1.7062	1.7602	1.6926	1.7741	1.6788	1.7881	1.6649	1.8024
151	1.7207	1.7473	1.7072	1.7609	1.6937	1.7747	1.6800	1.7886	1.6662	1.8028
152	1.7216	1.7481	1.7083	1.7616	1.6948	1.7752	1.6812	1.7891	1.6675	1.8032
153	1.7226	1.7488	1.7093	1.7622	1.6959	1.7758	1.6824	1.7896	1.6688	1.8036
154	1.7235	1.7496	1.7103	1.7629	1.6971	1.7764	1.6836	1.7901	1.6701	1.8040
155	1.7244	1.7504	1.7114	1.7636	1.6982	1.7770	1.6848	1.7906	1.6714	1.8044
156	1.7253	1.7511	1.7123	1.7642	1.6992	1.7776	1.6860	1.7911	1.6727	1.8048
157	1.7262	1.7519	1.7133	1.7649	1.7003	1.7781	1.6872	1.7915	1.6739	1.8052
158	1.7271	1.7526	1.7143	1.7656	1.7014	1.7787	1.6883	1.7920	1.6751	1.8055
159	1.7280	1.7533	1.7153	1.7662	1.7024	1.7792	1.6895	1.7925	1.6764	1.8059
160	1.7289	1.7541	1.7163	1.7668	1.7035	1.7798	1.6906	1.7930	1.6776	1.8063
161	1.7298	1.7548	1.7172	1.7675	1.7045	1.7804	1.6917	1.7934	1.6788	1.8067
162	1.7306	1.7555	1.7182	1.7681	1.7055	1.7809	1.6928	1.7939	1.6800	1.8070
163	1.7315	1.7562	1.7191	1.7687	1.7066	1.7814	1.6939	1.7943	1.6811	1.8074
164	1.7324	1.7569	1.7200	1.7693	1.7075	1.7820	1.6950	1.7948	1.6823	1.8078
165	1.7332	1.7576	1.7209	1.7700	1.7085	1.7825	1.6960	1.7953	1.6834	1.8082
166	1.7340	1.7582	1.7218	1.7706	1.7095	1.7831	1.6971	1.7957	1.6846	1.8085
167	1.7348	1.7589	1.7227	1.7712	1.7105	1.7836	1.6982	1.7961	1.6857	1.8089
168	1.7357	1.7596	1.7236	1.7718	1.7115	1.7841	1.6992	1.7966	1.6868	1.8092
169	1.7365	1.7603	1.7245	1.7724	1.7124	1.7846	1.7002	1.7970	1.6879	1.8096
170	1.7373	1.7609	1.7254	1.7730	1.7134	1.7851	1.7012	1.7975	1.6890	1.8100
171	1.7381	1.7616	1.7262	1.7735	1.7143	1.7856	1.7023	1.7979	1.6901	1.8103
172	1.7389	1.7622	1.7271	1.7741	1.7152	1.7861	1.7033	1.7983	1.6912	1.8107
173	1.7396	1.7629	1.7279	1.7747	1.7162	1.7866	1.7042	1.7988	1.6922	1.8110
174	1.7404	1.7635	1.7288	1.7753	1.7171	1.7872	1.7052	1.7992	1.6933	1.8114
175	1.7412	1.7642	1.7296	1.7758	1.7180	1.7877	1.7062	1.7996	1.6943	1.8117
176	1.7420	1.7648	1.7305	1.7764	1.7189	1.7881	1.7072	1.8000	1.6954	1.8121
177	1.7427	1.7654	1.7313	1.7769	1.7197	1.7886	1.7081	1.8005	1.6964	1.8124
178	1.7435	1.7660	1.7321	1.7775	1.7206	1.7891	1.7091	1.8009	1.6974	1.8128
179	1.7442	1.7667	1.7329	1.7780	1.7215	1.7896	1.7100	1.8013	1.6984	1.8131
180	1.7449	1.7673	1.7337	1.7786	1.7224	1.7901	1.7109	1.8017	1.6994	1.8135
181	1.7457	1.7679	1.7345	1.7791	1.7232	1.7906	1.7118	1.8021	1.7004	1.8138

182	1.7464	1.7685	1.7353	1.7797	1.7241	1.7910	1.7128	1.8025	1.7014	1.8141
183	1.7471	1.7691	1.7360	1.7802	1.7249	1.7915	1.7137	1.8029	1.7023	1.8145
184	1.7478	1.7697	1.7368	1.7807	1.7257	1.7920	1.7146	1.8033	1.7033	1.8148
185	1.7485	1.7702	1.7376	1.7813	1.7266	1.7924	1.7155	1.8037	1.7042	1.8151
186	1.7492	1.7708	1.7384	1.7818	1.7274	1.7929	1.7163	1.8041	1.7052	1.8155
187	1.7499	1.7714	1.7391	1.7823	1.7282	1.7933	1.7172	1.8045	1.7061	1.8158
188	1.7506	1.7720	1.7398	1.7828	1.7290	1.7938	1.7181	1.8049	1.7070	1.8161
189	1.7513	1.7725	1.7406	1.7833	1.7298	1.7942	1.7189	1.8053	1.7080	1.8165
190	1.7520	1.7731	1.7413	1.7838	1.7306	1.7947	1.7198	1.8057	1.7089	1.8168
191	1.7526	1.7737	1.7420	1.7843	1.7314	1.7951	1.7206	1.8061	1.7098	1.8171
192	1.7533	1.7742	1.7428	1.7848	1.7322	1.7956	1.7215	1.8064	1.7107	1.8174
193	1.7540	1.7748	1.7435	1.7853	1.7329	1.7960	1.7223	1.8068	1.7116	1.8178
194	1.7546	1.7753	1.7442	1.7858	1.7337	1.7965	1.7231	1.8072	1.7124	1.8181
195	1.7553	1.7759	1.7449	1.7863	1.7345	1.7969	1.7239	1.8076	1.7133	1.8184
196	1.7559	1.7764	1.7456	1.7868	1.7352	1.7973	1.7247	1.8079	1.7142	1.8187
197	1.7566	1.7769	1.7463	1.7873	1.7360	1.7977	1.7255	1.8083	1.7150	1.8190
198	1.7572	1.7775	1.7470	1.7878	1.7367	1.7982	1.7263	1.8087	1.7159	1.8193
199	1.7578	1.7780	1.7477	1.7882	1.7374	1.7986	1.7271	1.8091	1.7167	1.8196
200	1.7584	1.7785	1.7483	1.7887	1.7382	1.7990	1.7279	1.8094	1.7176	1.8199