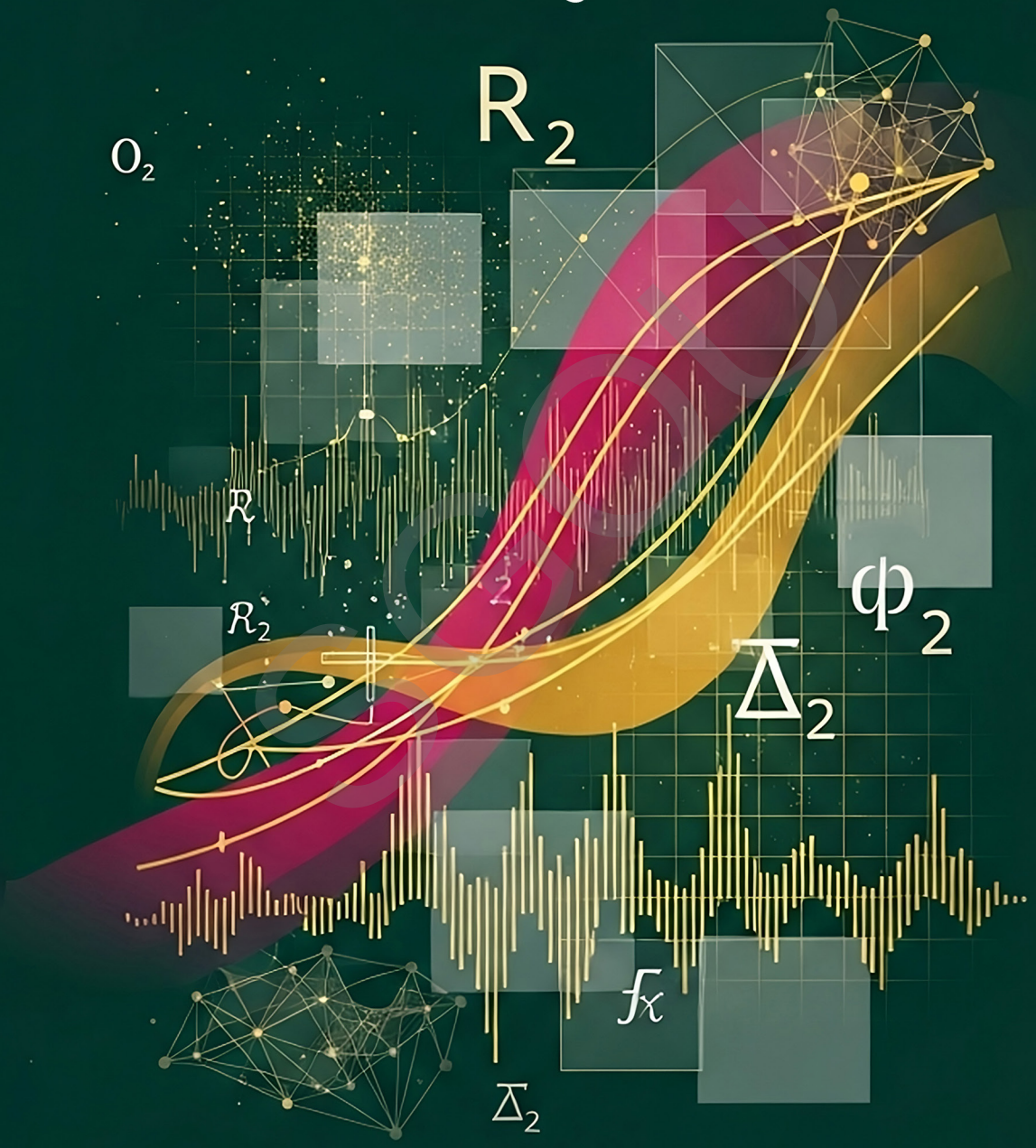
ECONOMETRICS

COURSE CODE: M23EC11DC

Postgraduate Programme in Economics

Discipline Core Course

Self Learning Material



SREENARAYANAGURU OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU OPEN UNIVERSITY

Vision

To increase access of potential learners of all categories to higher education, research and training, and ensure equity through delivery of high quality processes and outcomes fostering inclusive educational empowerment for social advancement.

Mission

To be benchmarked as a model for conservation and dissemination of knowledge and skill on blended and virtual mode in education, training and research for normal, continuing, and adult learners.

Pathway

Access and Quality define Equity.

Econometrics

Course Code: M23EC11DC

Semester - IV

Discipline Core Course
Postgraduate Programme in Economics
Self Learning Material
(With Model Question Paper Sets)

SREENARAYANAGURU
OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

SREENARAYANAGURU
OPEN UNIVERSITY

ECONOMETRICS

Course Code: M23EC11DC
Semester- IV
Discipline Core Course
Postgraduate Programme in Economics

Scan this QR Code for reading the SLM on a digital device.

Edition
October 2025

ISBN 978-81-990500-9-9

9 788199 050099

www.sgou.ac.in

Visit and Subscribe our Social Media Platforms

MESSAGE FROM VICE CHANCELLOR

Dear learner,

I extend my heartfelt greetings and profound enthusiasm as I warmly welcome you to Sreenarayanaguru Open University. Established in September 2020 as a state-led endeavour to promote higher education through open and distance learning modes, our institution was shaped by the guiding principle that access and quality are the cornerstones of equity. We have firmly resolved to uphold the highest standards of education, setting the benchmark and charting the course.

The courses offered by the Sreenarayanaguru Open University aim to strike a quality balance, ensuring students are equipped for both personal growth and professional excellence. The University embraces the widely acclaimed "blended format," a practical framework that harmoniously integrates Self-Learning Materials, Classroom Counseling, and Virtual modes, fostering a dynamic and enriching experience for both learners and instructors.

The University aims to offer you an engaging and thought-provoking educational journey. The postgraduate programme in Economics builds on the undergraduate programme by covering more advanced theories and practical applications. The course material aims to spark learners' interest by using real-life examples and combining academic content with empirical evidence, making it relevant and unique. The Self-Learning Material has been meticulously crafted, incorporating relevant examples to facilitate better comprehension.

Rest assured, the university's student support services will be at your disposal throughout your academic journey, readily available to address any concerns or grievances you may encounter. We encourage you to reach out to us freely regarding any matter about your academic programme. It is our sincere wish that you achieve the utmost success.

Regards,
Dr. Jagathy Raj V.P.

01-10-2025

Contents

BLOCK 1

Introduction to Econometrics

UNIT 1

Scope, Methodology and Limitations of Econometrics

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the basics of theoretical and applied econometrics
- explain how econometrics integrates economic theory, mathematics, and statistical inference
- design an empirically testable economic model, articulating underlying economic theory and testable hypotheses
- critically discuss common pitfalls—model misspecification, omitted-variable bias, measurement error, sample-selection bias
- explain identification challenges and the limits of causal inference in observational data

Background

Economic theories are formulated by depicting factual evidences with the help of mathematical and statistical observations. In this efforts, theoretical and applied econometrics integrates economic theory, mathematics and statistical inference. It also guides in formulating testable hypotheses and estimation of results with reliability and validity. This unit unfolds the meaning of econometrics, stating its scope and methodology and helps in acquiring knowledge regarding:

- The role of statistics in economics.
- How mathematical models provide structure to economic ideas.
- The importance of quantitative, data-driven analysis in modern economic research.

Keywords

Mathematical Economics, Statistics, Empirical Validation, Policy Making, Forecasting, Regression Analysis, Hypothesis Testing, Time Series Data, Cross-Section Data, Panel Data, Causality

Discussion

The term econometrics was coined in 1926 by Ragnar A. K.Frisch, a Norwegian economist who shared the first Nobel Prize in Economics in 1969 with Jan Tinbergen. Many economists had used data and made calculations long before 1926, but it was Frisch who felt the significance of a new term associated with the interpretation and use of data in Economics. Today, Econometrics has evolved as a broad area of study within economics incorporating timebound changes according to the emergence of new tools and techniques.

Econometrics is concerned with the measurement of economic relationships through an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships. The relationships of economic theories are expressed in mathematical forms, combined with empirical economics. The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships. These relationships depict the random behaviour of economic concepts which are generally not considered in economics and mathematical formulations.

1.1.1 Definitions

1. Econometrics may be defined as the quantitative analysis of actual economic phenomenon based on the concurrent development of theory and observation, related by appropriate methods of inference. (P.A. Samuelson, T.C.Koopman, J.R.N Stone)

2. Econometrics is concerned with the empirical determination of economic laws (H.Theil)

3. Econometrics may be defined as the social science in which the tools of economic theory, mathematics and statistical inference are applied to the analysis of economic phenomena (Arthur S.Goldberg)

4. Econometrics consists of the application of mathematical statistics to economic data to lend empirical support to the model constructed on mathematical economics and to obtain numerical results. (Gerhard. Tinter)

5. Every application of mathematics or of statistical methods to the study of economic phenomena (Malinvaud 1966)

6. The production of quantitative economic statements that either explain the behaviour of variables we have already seen, or forecast (ie. predict) behaviour that we have not yet see, or both (Christ 1966)

7. Econometric is the art and science of using statistical methods for the measurement of economic relations (Chow, 1983).

1.1.2 Scope of Econometrics

Goals/Scope of Econometrics means the importance or usefulness of the science, Econometrics. There are mainly three main goals for the subject. They are,

- Analysis,
- Policy making, and
- Forecasting

1. **Analysis : Testing Economic Theory :** Earlier economic theories started with a set of assumptions concerning the behaviour of some individual units (consumers, producers etc.). From these assumptions the economists derived some general conclusions or laws determining the working process of the economic system. Economic theories thus developed in an abstract level were not tested against economic reality. In other words, attempts were not made to examine whether the theories explained the actual economic behaviour of individuals. An empirical validation of the theories is being provided

   Econometrics primarily aims at the verification of economic theories. Under such circumstances we can say that the purpose of the research is analysis, i.e., to obtain the empirical evidences to test the explanatory power of economic theories and to decide how well they explain the observed behaviour of economic units. Today, any theory, regardless of its elegance in exposition or its sound logical consistency, cannot be established and generally accepted without some empirical testing.

2. **Policy Making** : Obtaining numerical estimates of the coefficients of economic relationships for policy simulations. In many cases we apply the various econometric techniques in order to obtain reliable estimates of the individual coefficients of economic relationships from which we can evaluate elasticity or other parameters of economic theory. The knowledge of obtaining the numerical value of (For example, the Marginal concepts in Economics, Concept of multiplier, technical coefficients of production, level of investment etc) coefficients is very important for the formulation of the economic policies of the governments. It helps to compare the effects of various alternative policy decisions.

   For example, the decision of the government about devaluing the currency will depend to a great extent on the numerical values of Marginal propensities of imports and exports and as well as the numerical values of price elasticities of imports and exports (ei and ex). If the sum of the price elasticities of imports and exports is less than one (ei + ex <1) in absolute value, the devaluation will not help in eliminating the deficit in BOP. This shows that how important is the numerical value of the coefficients of economic relationships. Econometrics can provide such numerical estimates and has become an essential tool for the formulation of sound economic policies.

3. **Forecasting the Future Values of Economic Magnitudes :** Forecasting the values of economic variables is essential in framing different economic policies and econometrics will help a lot for such forecasting and policy framing. For example, suppose government is going to frame its poverty policy, under such circumstances it is necessary to know what the current employment situation is, what will be the level of poverty in the next five years if the Government doesn't take any apt anti -poverty programmes. The facts and figures gained through such estimates will help the government to deal with different situations like: If poverty in the future is high, government should take appropriate measures to reduce it.

   Forecasting is thus becoming increasingly important for the regulation of developed economies as well as for the planning of the economic development for the underdeveloped countries. Simple econometric tools such as trend analysis, regression models, and time-series forecasting can be used to make such predictions.

1.1.3 Methodology of Econometrics

How does the econometrician go ahead in analysing an economic theory? There came the role of methodology in econometrics, it is in fact a step-by-step procedure. These steps are:

1. **Statement of the theory/hypothesis** : A theory should have a prediction. In statistics and econometrics, we also speak of hypothesis. Hypothesis is an if-then Proposition and the theory is in fact a validated hypothesis. One example is about the value of the Marginal Propensity to Consume (MPC) proposed by Keynes, 0<MPC<1. Other examples could be that lower taxes would increase growth, or maybe that it would increase economic inequality, and that introducing a common currency has a positive effect on trade.

2. **Specification of mathematical model** : A model is a simplified representation of a real-world. It should be representative in the sense that it should contain the salient features of the phenomena understudy. In general, one of the objectives in modelling is to have a simple model to explain a complex phenomenon. Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic. This is where the algebra enters. We need to use mathematical skills to generate an equation. Assume a theory predicting that more schooling increases the wage, ie, a positive relation between years of schooling and wage rate. In economic terms, we say that the return to schooling is positive on wage. The equation is: $Y = \beta_0 + \beta_1 X$ where; Y, the dependent variable as the variable for wage and β0 is a constant and β_1 is the coefficient of schooling, and X, the independent variable is a measurement of schooling, i.e. the number of years in school. We also call β0 as intercept and β1 as the slope coefficient. Normally, we would expect both β0 and β1 to be positive.

3. **Specification of Econometric model** : An economic model is a set of assumptions that describes the behaviour of an economy, or more generally, a phenomenon.

In practice, generally, all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model. Rest of the variables are dumped in a basket labelled—disturbances-- where the disturbances are random variables whose behaviour is unpredictable.

This is the main difference between economic modelling and econometric modelling. The mathematical modelling is exact in nature, whereas the econometric modelling contains a stochastic term also.

Here, we assume that the mathematical model is correct but we need to account for the fact that it may not be so. We add an error term, 'u' to the equation above. It is also called a random variable or stochastic variable. It represents other non-quantifiable or unknown factors that affect Y, but are not explicitly included in the model. It also represents errors of measurements that may have entered the data. The econometric equation is:

$$\beta_0 + \beta_1 X + U$$

The error term, U, is assumed to follow some sort of statistical distribution.

4. **Collection of data** : We need data for the variables above. This can be obtained from government statistics agencies and other sources. A lot of data can also be collected on the Internet in these days. But we need to learn the art of finding appropriate data from the ever-increasing heaps of data. Data could be from primary or secondary sources.

5. **Estimation of the model** : Here, we quantify β1 and β2 i.e. we obtain numerical estimates. This is done by statistical technique called regression analysis. This provides the empirical content to the theory under consideration.

6. **Hypothesis testing** : In this stage of testing the hypothesis first we have to consider the theory and the hypothesis that we explained in earlier stages. The prediction of our theory was that schooling is good for the wage. Does the econometric model support this hypothesis? What we do here is called statistical inference (hypothesis testing). Technically speaking, to have this positive relationship between years of schooling and wages, the β2 coefficient of the econometric model should be greater than 0.

The iterative process can be depicted in the following figure:

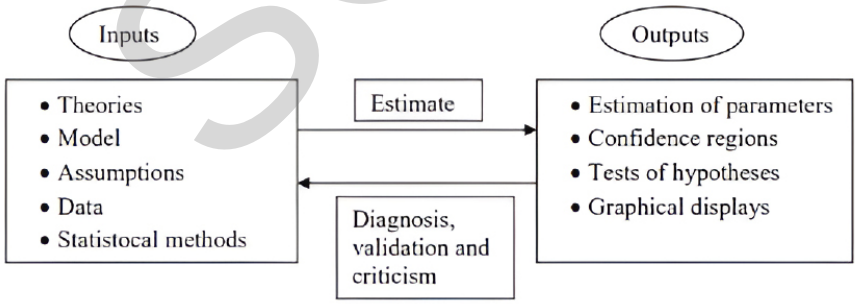


Fig. 1.1.1 Iterative Process

7. **Forecasting/ prediction** : If the hypothesis testing was positive, i.e. the theory was concluded to be correct; we forecast the values of the wage by predicting the values of education. For example, how much would someone earn for an additional year of schooling? If the X variable is the years of schooling, the β_2 coefficient gives the answer to the question.

8. **Use of the model for policy purpose** : Lastly, if the theory seems to make sense and the econometric model was not refuted on the basis of the hypothesis test, we can go on to use the theory for policy recommendation.

1.1.4 Nature and Sources of Data for Econometric Analysis

Various types of data are used in the estimation of the model.

1. **Time series data** : Time series data give information about the numerical values of variables from collected over a period of time. For example, the data during the years 2010 – 2020 for monthly income constitutes time series data.

2. **Cross-section data** : The cross-section data give information on the variables concerning individual agents (e.g., consumers or produces) at a given point of time. For example, data collected from a sample of consumers and their family budgets showing expenditures on various commodities by each family, as well as information on family income, family composition and other demographic, social or financial characteristics is an example of cross section data, as we collect all these information at a point of time

3. **Panel data** : The panel data are the data from a repeated survey of a single (cross-section) sample in different periods of time.

4. **Pooled Data** : Pooled data refers to data that combines cross-sectional and time-series observations. It means observations on multiple units (like individuals, firms, or countries) are collected over multiple time periods, and then pooled together to estimate a common regression model. Therefore the pooled dataset as a whole has both cross-sectional and time-series features. Each sample is a new, random draw from the same population in a different time period.

1.1.5 Contribution of Econometric Methods to the Development of Economics

Econometric methods have significantly contributed for the evolving of economics discipline by enabling the empirical testing of economic theories, quantifying relationships between variables, and improving forecasting and policy evaluation. They bridge the gap between theoretical models and real-world data, providing an evidence-based approach to understanding and addressing economic issues.

1. **Empirical Validation of Economic Theories** : Econometrics provides the tools to test whether the relationships suggested by economic theory actually hold true in the real world. For example, econometric methods can be used to verify the relationship between inflation and unemployment, or the impact of interest rates on investment. By analysing historical data, econometricians can either confirm or reject economic models, thus refining our understanding of economic phenomena.

2. **Quantification of Economic Relationships** : Econometrics allows economists to move beyond qualitative statements (e.g., "higher income leads to more spending") and quantify these relationships (e.g., "for every dollar increase in income, spending increases by 95 cents"). This quantification is crucial for building accurate economic models and making informed predictions.

3. **Forecasting and Policy Evaluation** : Econometric models are widely used for forecasting future economic trends and evaluating the potential impact of policy changes. For instance, policymakers can use econometric models to simulate the effects of tax cuts or

interest rate adjustments before implementing them. This allows for a more data-driven approach to policymaking, potentially leading to more effective and efficient economic management.

4. **Understanding Causality** : Econometrics helps in establishing causal relationships between variables, which is crucial for understanding the underlying mechanisms driving economic outcomes. For example, econometric methods can be used to determine whether a specific policy intervention actually caused a change in economic growth.

5. **Advancements in Various Economic Fields** : Econometric methods are applied across various fields of economics, including macroeconomics, microeconomics, finance, and labour economics. The development of econometric techniques has been instrumental in the progress of these fields, allowing for a deeper understanding of complex economic phenomena.

In summary, econometric methods have transformed economics from a largely theoretical discipline into an empirical science, providing the tools to test theories, quantify relationships, make predictions, and evaluate policies based on evidence. This has led to a more refined and data-driven approach to understanding and addressing economic issues, benefiting both academic research and practical policymaking.

1.1.6 Limitations of Econometrics

Like any other subjects, econometrics also not free from limitations. Some of them are;

a. It is concerned only with quantifiable phenomena like prices, production, employment etc. It throws very little light on qualitative problems.

b. All the econometric analysis is based on data availability. The available data may be insufficient and inaccurate.

c. Predictions are made through sampling methods. Therefore, the limitations of the sampling method are also categorized as the limitations of econometrics.

d. The statistical methods used in econometrics are based on certain assumptions, which are not true with economic data.

e. Econometric methods are time consuming, tedious and complex. It requires a sound knowledge of mathematics and statistics.

Summarised Overview

Econometrics, a term introduced by Ragnar A.K. Frisch in 1926, has grown into a specialised branch of economics that integrates economic theory, mathematics, and statistics to measure and interpret economic relationships. It goes beyond traditional theoretical models by quantifying parameters, accounting for randomness in behaviour, and providing empirical validation. Definitions by several economists highlight econometrics as a discipline centred on quantitative analysis, empirical determination of laws, and the application of statistical methods to support and test economic models. The scope of econometrics can be broadly categorised into three goals: analysis, policy-making, and forecasting. In analysis, econometrics tests economic theories against real-world data to verify their explanatory power. In policy-making, it provides numerical estimates of coefficients in economic relationships, allowing policymakers to simulate outcomes and make informed decisions, such as evaluating the impact of devaluation or taxation. Forecasting involves predicting future economic variables, such as poverty levels or employment, using econometric models to inform both developed and developing economies. The methodology of econometrics follows systematic steps: formulating hypotheses, specifying mathematical and econometric models, collecting data, estimating parameters (using techniques like regression analysis), conducting hypothesis testing, forecasting outcomes, and applying results for policy purposes. The models often incorporate error terms to account for unpredictable disturbances and measurement errors.

Econometrics employs different types of data: time series data (collected over periods of time), cross-sectional data (collected at a point in time across agents), panel data (repeated observations of the same units), and pooled data (qualitative data represented through dummy variables). The contribution of econometrics to economics has been profound. It allows empirical validation of theories, quantification of economic relationships, forecasting of trends, policy evaluation, and understanding of causality. These contributions extend across macroeconomics, microeconomics, finance, and labour economics, transforming economics into a more empirical and data-driven science. However, limitations remain. Econometrics deals mainly with quantifiable phenomena, leaving out qualitative aspects. Its effectiveness is constrained by the availability and reliability of data, the limitations of sampling methods, and the restrictive assumptions of statistical techniques. Moreover, econometric analysis is complex, time-intensive, and demands strong mathematical and statistical skills. In essence, econometrics bridges theory and practice by offering tools to validate models, guide policy, and predict future trends, while also facing constraints rooted in data quality and methodological assumptions.

Assignments

1. "Econometrics is the unification of economic theory, mathematics and statistical inference." Explain this statement and contrast it with the roles of mathematical economics and economic statistics.
2. Enumerate and discuss each step in a standard econometric research design, from model formulation to policy evaluation. Illustrate with a working example.
3. In the era of machine learning, do traditional econometric identification concerns still matter? Argue for or against, citing examples.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*): Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 2

Regression Functions and Functional Forms

Learning Outcomes

After completing this unit, the learner will be able to:

- understand analytically the difference between population regression function and sample regression function
- state, justify, and test each Gauss–Markov assumption
- compare precision and bias of origin-restricted versus unrestricted models in econometrics
- understand alternative functional forms

Background

Key concepts in regression analysis like PRF and SRF can be analysed in detail specifying their significance in analysis. To ensure valid and unbiased estimation, several assumptions must be satisfied while using a simple linear regression model. These assumptions can be studied and an in-depth observation can be made in this unit.

Keywords

Population Regression Function, Error Term, Sample Regression Function, Ordinary Least Squares, Maximum Likelihood, Simple Linear Regression, Log–Log Model, Elasticity, Log–Lin Model, Growth Rate, Lin–Log Model, Reciprocal Model

Discussion

Some important terms and terminologies such as population regression function, sample regression function, significance of stochastic disturbance term etc. that are frequently used in the analysis of regression models. Let us have a look at them.

1.2.1 Population Regression Function (PRF)

The group of individuals or items under study is known as the population. In statistics, population is the aggregate of facts or objects, animate or inanimate, under study in any statistical investigation. Informally, it means the set of all possible outcomes of an experiment or measurement. We always expect an idealistic situation which is possible only with a population. This is a highly idealistic situation and very rare to occur. Even the distinction between population and sample is relative and such a distinction is necessary in economic studies.

A Population Regression Function (PRF) can be defined as the average value of the dependant variable for a given value of the independent variable. In other words, PRF tries to find out how the average value of the dependant variable varies with the given value of the explanatory variable.

To explain the PRF. We are taken an example of a hypothetical country with a total population of 50 families. Suppose we are interested in studying the relationship between weekly family expenditure (Y) and weekly disposable income (X). That is, we want to predict the (population) mean level of weekly consumption expenditure knowing the family's weekly income. Suppose we divide these 50 families into 10 groups of approximately same income and examine the consumption expenditure of families in each of these income groups. This hypothetical data can be shown as follows.

Table 1.2.1 Family Income Data

X Weekly family income ⇨	10	12	15	18	20	23	25	28	30	35
Y weekly family expenditure ⇩	5	5	8	11	11	15	16	16	20	22
	6	8	9	12	14	16	17	18	21	23
	7	9	10	13	15	17	18	20	23	24
	8	10	11	14	16	18	21	22	24	26
	9		13	15		19	23	24		30
			15			23		26		
Total	35	32	66	65	56	108	95	126	88	125

The above table can be interpreted as follows.

Corresponding to the weekly income 10, there are 5 families whose weekly consumption expenditure range between 5 and 9. Similarly, given X=30, there are 4 families whose weekly consumption expenditure falls between 20 and 24. In other words, each column of the table above gives the conditional distribution of 'Y'conditional upon the given values of X.

From this table we can easily compute conditional probabilities of y, P(Y/X), as follows:

For X=10, there are 5 y values, 5,6,7,8 and 9. Therefore, given X= 10, the probability of obtaining any of these consumption expenditure is 1/5. Symbolically,

P(Y=7/X=10) = 1/5. Or

P (Y =23/X=30) = ¼ and so on.

These conditional probabilities can be given in Table 1.2.2 for the values given in Table 1.2.1.

Table 1.2.2 Conditional Probability

X ⇨	10	12	15	18	20	23	25	28	30	35
P(y/Xi) ⇩ Conditional probabilities	1/5	¼	1/6	1/5	¼	1/6	1/5	1/6	¼	1/5
	1/5	¼	1/6	1/5	¼	1/6	1/5	1/6	¼	1/5
	1/5	¼	1/6	1/5	¼	1/6	1/5	1/6	¼	1/5
	1/5	¼	1/6	1/5	¼	1/6	1/5	1/6	¼	1/5
	1/5		1/6	1/5		1/6	1/5	1/6		1/5
			1/6			1/6		1/6		
Conditional Means of y	7	8	11	13	14	18	19	21	22	25

From the conditional probability distribution of Y we can compute its mean or average value, known as the, conditional mean or conditional expectation denoted by E(Y/X=Xi) and simply as E(Y/Xi).

When we plot the data of weekly family consumption expenditure at different levels of income on a graph paper, we get a scatter diagram as follows (Figure given below).

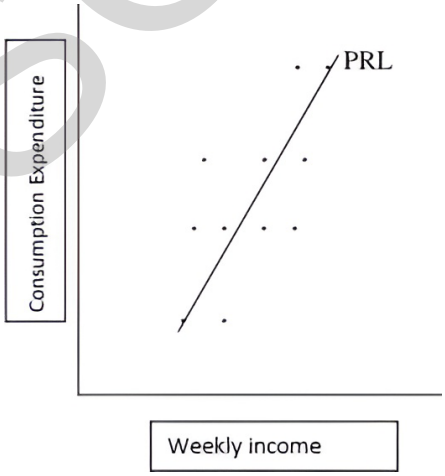


Fig. 1.2.1 Population Regression Line

The Figure 1 implies very clearly that consumption expenditure on an average basis increase as income increases. When we include the conditional means of Y, we get a straight line with a positive slope. This line is known as the Population Regression Line (PRL).

Geometrically, a population regression curve is simply the locus of the conditional means or expectations of the dependent variable for the fixed values of the independent variables. Therefore, it is clear that each conditional mean is a function of Xi.

Symbolically,

$$E(Y/Xi) = f(X_i). \text{ ---------- } (1)$$

The above equation is known as the Population Regression Function (PRF). It merely states that the population means of the distribution of Y, given X, is functionally related to Xi.

Taking our example of consumption function,

$$E(Y/Xi) = \beta_0+\beta_1X_i \text{----------} (2)$$

The above example is a linear population regression function.

Stochastic Specification of PRF

The PRF states that the mean or average responses of Y varies with X. Consider our consumption expenditure figures, at income level 20, it can be seen as 11, 14, 15 and 16 with a mean 14. The PRF indicates the average only. The deviation from mean of the actual expenditure figures are not explained by the PRF. Therefore, when we take one consumer at random, his consumption expenditure may be greater or less than the mean value. So this can be expressed by the stochastic specification of PRF as;

Yi = E(Y/Xi) + ui Where; ui = Stochastic error term.

ui may be defined as an unobservable random variable taking positive or negative values. It is also termed as Stochastic disturbance term.

Therefore, we can explain the stochastic PRF as; the expenditure of a family, given its income level as the sum of two components.

1.E(Y/Xi) - conditional mean expenditure and

2. u_i - random component.

Therefore; our estimated consumption function can be expressed as;

$Yi = E(Y/X_i) + u_i$ or

$Yi = \beta_0+\beta_1X_i + u_i$ ---------- -------(3)

The stochastic specified PRF clearly shows that there are other variables besides income that affect consumption expenditure and that an individual family's consumption expenditure cannot be fully explained by the regression model.

1.2.2 Sample Regression Function (SRF)

Practically, it is not possible to rely on population studies always. Under such circumstances we have to rely on sample studies associated with this we face sampling related problems too.

Therefore, our task is to estimate the PRF on the basis of the sample information.

For this, we randomly select some of the Y values corresponding to fixed values of X from the given population. In this way, we have to draw so many samples from the population. Considering our example, we can draw a sample from the given population of income and consumption expenditure as follows (Table 1.2.3.).

The Table 1.2.3 is a sample of consumption expenditure at different levels of income. Like this we have to draw 'n' different samples from the population.

When we estimate the average value of the dependent variable with the help of a sample it is called stochastic Sample Regression Function (SRF) to estimate the PRF. Actually, to estimate the numerical values of βs in PRF, we have to depend on the whole data. But in practice, we are interested in a sample and with the help of the sample, we are trying to estimate the PRF.

When the plot the sample data on consumption expenditure on a graph paper we have the Figure given below

Table 1.2.3 Sample Data on Consumption Expenditure

Weekly Family Income	Weekly Family consumption Expenditure
10	7
12	8
15	11
18	13
20	15
23	17
25	18
28	20
30	21
35	24

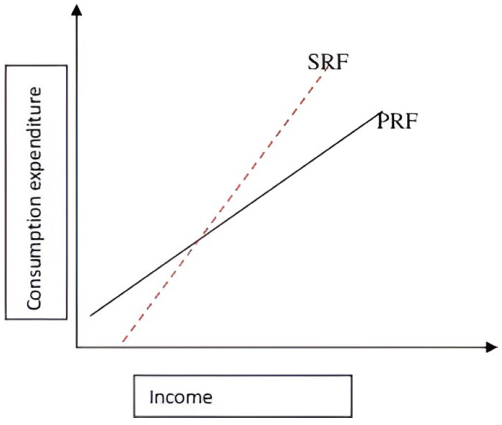


Fig. 1.2.2. Sample Regression Line (SRL)

The SRF can be expressed as;

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \quad \ldots\ldots (4)$$

Then our objective is to estimate the PRF, $Yi = \beta_0 + \beta_1 X_i + u_i$ on the basis of the SRF,

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$.

Here the SRF is the estimator of the PRF. We are using SRF to find PRF. That is, $\hat{\beta}_0$ as the estimator of β_0, $\hat{\beta}_1$ as the estimator of β_1 and ûi as the estimator of Ui. An estimator is a rule, formula, or method used to calculate an estimate of an unknown population parameter based on sample data.

The graphical representation of SRF is termed as sample regression line SRL. To conclude, we can say that the primary objective of regression analysis is to estimate PRF on the basis of the SRF. We may have to select as many samples as possible to reduce the sampling fluctuations, so that it will become easier to approximate the SRF to the PRF.

So in the linear regression analysis we are trying to estimate the average value of the dependent variable Y for any given values of the independent variable Xi. For this we are estimating β_0 and β_1 with the help of $\hat{\beta}_0$ and $\hat{\beta}_1$. For the estimation of a linear regression model there are mainly two methods;

1. Ordinary Least Square (OLS) method, and
2. Maximum Likelihood (ML) method

The method of OLS is extensively used in regression analysis because of its mathematical simplicity. But both methods provide the same results.

1.2.3 Significance of Error term/ Stochastic Disturbance term (u_i)

The significance of incorporating the term u_i in the econometrics model is due to the following reasons:

1. **Vagueness of theory :** Sometimes we may ignorant or unsure about the other variables affecting Y (consumption) other than income. Under such circumstances, the theory may be incomplete to explain the behaviour of Y. Therefore, ui may be used as a substitute for all the excluded or omitted variables from the model.

2. **Unavailability of data :** Even if we know all other variables affecting Y other than X, there may not have quantitative information about these variables. Therefore, we are forced to omit some variables from our model despite its great theoretical relevance. Hence ui may be implied these omitted variables.

3. **Core variables versus Peripheral variables :** Apart from the influence of different variables, there may be some variables which jointly influence the model, which is not explicitly in the model. ui thus tries to explain the combined effect of these variables in the model concerned.

4. **Intrinsic randomness in human behaviour :** Even if we succeed in introducing all the relevant variables into the model there may be some intrinsic randomness in the human behaviour. Therefore, ui may well reflect this intrinsic randomness.

5. **Poor proxy variables** : Although the classical regression model assumes that the variables Y and X are measured accurately, in practice, the data may be plagued by errors of measurement. But since data on these variables are not directly observable in practice, we may use proxy variables. The disturbance term ui may in this case also represent the errors of measurement.

6. **Principle of parsimony** : By this principle, we would like to keep our regression model as simple as possible. It is done by avoiding some variables from the model. Let ui represents all the omitted variables.

7. **Wrong functional form :** Even if we have theoretically correct variables explaining a phenomenon, but unfortunately due to the unavailability of data on these variables, we do not know the form of the functional relationship between the regressand and the regressors. In two variable models, the functional form of the relationship can often be gained from the scatter diagram but in a multiple regression model it is not easy to determine the appropriate functional form graphically. So, the errors of the functional form can also be solved by the inclusion of the random variable ui.

1.2.4 Simple Linear Regression Model

One of the very important roles of econometrics is to provide the tools for modelling on the basis of given data. The regression modelling technique helps a lot in this task. The regression models can be either linear or non-linear based on which we have linear regression analysis and non-linear regression analysis.

1.2.4.1 Linear Regression Model

The term 'Regression' was introduced by Francis Galton and Galton's Law of Universal regression was confirmed by his Friend, Karl Pearson. The modern interpretation of regression is quite different from their analysis.

Regression analysis is concerned with the study of the dependence of one variable (dependant variable), on one or more other variables, the explanatory variables (Independent Variable), with a view to estimating and/or predicting the mean or average value of the former in terms of the known or fixed values of the latter.

The major objectives of regression analysis are as follows:

- To estimate the mean value of the dependant variable given the value of the independent variables.
- To test the hypothesis suggested by the underlying economic theory about the nature of the dependence.

- To predict or forecast the mean value of the dependant variable, given the values of the independent variables.

That is, through the regression analysis, we are going to estimate the mean value, or average value or expected value of the dependant variable 'Y' based on the known values of the independent variables X's. That is we are estimating

E(Y/Xi) ie, conditional expectation of Y given Xi.

Suppose the outcome of any process is denoted by a random variable Y, called as dependent (or study) variable, depends on ' k' independent (or explanatory) variables denoted by X_1, X_2, X_3.....X_k . Suppose the behaviour of Y can be explained by a relationship given by,

$Y = f(X_1, X_2,..., X_k , \beta_1, \beta_2, ---- \beta_k)+u$

where f is some well-defined function and β_1, β_2, ------- β_k are the parameters which characterize the role and contribution of X_1, X_2, -------- X_k respectively. The term 'u' reflects the stochastic nature of the relationship between Y and X_1, X_2, ------ X_k and indicates that such a relationship is not exact in nature. When u=0, then the relationship is called the mathematical model otherwise the statistical model.

Here we are discussing a simple linear regression model. The term is broadly used to represent any phenomenon in a mathematical framework. To explain a simple linear model, two terms 'simple' and 'linear' must be explained first.

The term simple regression means, it is a regression in which the dependant variable is related to a single explanatory variable. It represents a fundamental idea of the regression analysis that a model must be as simple as possible. We have the multiple regressions also in which the regressand (Dependent Variable) is related to more than one regressors (Independent Variables). Therefore, in a simple regression model there are only two variables;

- One explained variable, and
- One explanatory variable.

For example, in the Keynesian theory of consumption, we are trying to analyse the relationship of consumption with the household income. Here consumption is the regressand and the household income is the regressor. This type of analysis is called simple regression or two variable regression analysis. This simple form can be expressed as;

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

A model or relationship is termed as linear if it is linear in parameters and non-linear, if it is not linear in parameters. In other words, if all the partial derivatives of Y with respect to each of the parameters β_1, β_2, ... β_k are independent of the parameters, then the model is called as a linear model. If any of the partial derivatives of Y with respect to any of the β_1, β_2, ---- β_k is not independent of the parameters, the model is called non-linear. Note that the linearity or non-linearity of the model is not described by the linearity or non-linearity of explanatory variables in the model.

For example,

$$y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \ log \ X_3 + \varepsilon$$

is a linear model because $\partial y / \partial \beta_i, (i = 1,2,3)$ are independent of the parameters β_i, (i=1,2,3,...). On the other hand,

$$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \ log \ X + \varepsilon$$

is a non-linear model because $\partial y / \partial \beta_1 = 2\beta_1 X_1$ depends on β_1, although $\partial y / \partial \beta_2$ and $\partial y / \partial \beta_3$ are independent of any of the β_1, β_2 or β_3.

When the function *f* is linear in parameters, then $Y = f(X_1, X_2..., X_k, \beta_1, \beta_2, \text{-----}\beta_k) + u$ is called a linear model and when the function f is non-linear in parameters, then it is called a non- linear model. In general, the function f is chosen as

$Y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ... \beta_k) = \beta_1 X_1 + \beta_2 X_2 + \ + \beta_k X_k$

to describe a linear model. Since X_1, X_2, ..., Xk are pre- determined variables and Y is the outcome, so both are known. Thus the knowledge of the model depends on the knowledge of the parameters β_1, β_2, ---- β_k. The statistical linear modelling essentially consists of developing approaches and tools to determine β_1, β_2, ... β_k in the linear mode $Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$ given the observations on *Y* and X_1, X_2 ,. , X_k.

Therefore, simple linear regression model (SLRM) means that,

- There are only two variables; one dependant and one independent and
- The relation between dependant and independent variables are linear in parameters (may or may not be linear in variables).

Different statistical estimation procedures, e.g., method of maximum likelihood, the principle of least squares, method of moments etc. can be employed to estimate the parameters of the model. The method of maximum likelihood needs further knowledge of the distribution of Y whereas the method of moments and the principle of least squares do not need any knowledge about the distribution of Y. The regression analysis is a tool to determine the values of the parameters given the data on Y and X_1, X_2, ------ X_k.

1.2.5 Regression Through the Origin

There are occasions when the two variable PRF assumes the following form;

$Yi = \beta_2 Xi + ui$.....(1)

In this model, the intercepted term is absent or zero, hence the name regression through the origin.

For example, in Capital Asset Pricing Model (CAPM) of modern portfolio theory, the risk premium may be expressed as;

(ERi – rf) = βi (ERm – rf) .....(2)

Where;

ERi = Expected rate of return on security _i‘

ERm = Expected rate of return on the market portfolio

rf = Risk free rate of return

βi = Beta coefficient, a measure of systematic risk

If capital markets work efficiently, then capital asset pricing

model postulates that security i‘s expected risk premium (ERi– rf) is equal to that security’s β coefficient times the expected market risk premium(ERm – rf).

For empirical purposes, equation (2) can be expressed as;

Ri – rf = βi (Rm-rf) + ui .....(3) or

Ri – rf =αi+ βi (Rm-rf) + ui .....(4)

Equation (4) is known as the market model. If capital pricing model holds, αi is expected to be zero. This form of regression is known as regression through origin.

1.2.6 Functional Forms of Regression Models

The relationship between X and Y can take different functional forms, helping to capture various economic behaviours and elasticities:

Table 1.2.4 Model Types

Model Type	Formula	Interpretation/Use
Linear	$Y = \alpha + \beta X + u$	Each unit change in X leads to β units change in Y
Log-Log	$\log Y = \alpha + \beta \log X + u$	β shows % change in Y per % change in X (elasticity)
Log-Lin	$\log Y = \alpha + \beta X + u$	β is the % change in Y for 1 unit change in X
Lin-Log	$Y = \alpha + \beta \log X + u$	β is the change in Y for 1% change in X
Reciprocal	$Y = \alpha + \beta \frac{1}{X} + u$	Useful when increases in X have diminishing effects on Y

1.2.7 Extension of the Two Variable Linear Regression Models

The classical linear regression model requires the parameters must be linear and the variables may or may not be linear. But we consider only models that are variable in parameters as well as in the variables. The models that are linear in parameters but not necessarily in the variables are considered under the head Extension of the two variable linear regression models.

As an extension of two variable linear regressions we have mainly three models

1. The log linear model
2. Semi-log models

   -Log-lin models, and

   -Lin-log model,
3. Reciprocal models

In all these models we are transforming non-linear models which are not linear in variables to a linear model for simplicity. Moreover, we are familiarising regression through origin as a special case of simple linear regression model.

1.2.7.1 Log Linear Model and Measurement of Elasticity

In the case of log linear model, we are transforming an exponential regression model to a linear model. For this first of all we are considering an exponential model as;

$Y_1 = \beta_1 x_i^{\beta_2} e^{u_i}$ ----------(1)

This exponential model can also be expressed in terms of logarithm as;

$\ln (Y_i) = \ln (\beta_1) + \beta_2 \ln (X_i) + u_i \ln (e)$ ----------(2)

Here;

ln = natural logarithm whose base is 'e' Therefore the model becomes;

$\ln (Y_i) = \ln (\beta_1) + \beta_2 \ln (X_i) + u_i$

(where, loge e =1) Substitute α for ln (β1) we have,

$\ln (Y_i) = \alpha + \beta_2 \ln (X_i) + u_i$ ----------(3)

Then the model becomes linear in parameters. The linearity can be obtained by using logarithm and hence we can apply OLS, such models are called log-log or double log or log linear models.

If the assumptions of classical linear regression model are fulfilled, the parameters of equation (3) can be obtained by the method of OLS by substituting it as;

$Y_i^* = \alpha + \beta_2 X_i^* + u_i$

The OLS estimators $\hat{a}$ and $\hat{\beta}_2$ thus obtained will be BLUE of α and β_2 respectively.

Important feature of the log linear regression model is that the slope of the coefficient of the model β_2, measures the elasticity of Y with respect to X. That is, the percentage changes in Y for a given percentage change in X. Thus, if Y represents the quantity of a commodity demanded and X its unit price, β_2 measures the price elasticity of demand which have considerable importance in economics.

Two special features of Log linear model are;

- It is a constant elasticity model, and

$\hat{\beta}_2$ - Even though $\hat{a}$ and $\hat{\beta}_2$ an unbiased estimator of α and β_2, β_1= antilog ($\hat{a}$) is a biased estimator.

Constant elasticity model gives a constant total revenue change for a given percentage change in price regardless of the absolute level of price. The original model and its transformation into Log- Linear model can be differentiated as figure given below.

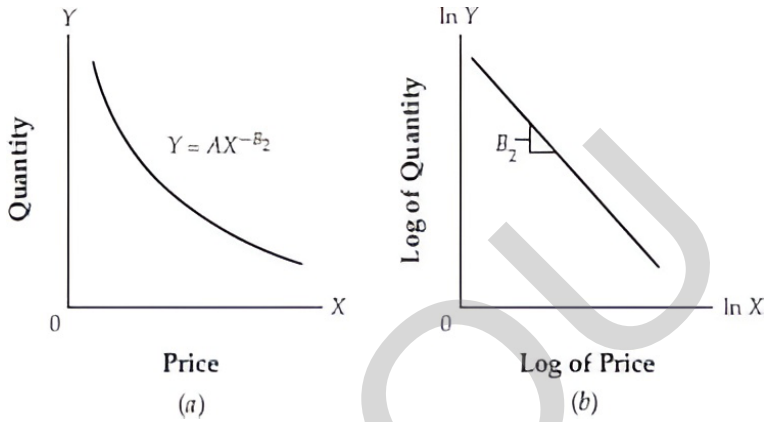


Fig. 1.2.3. Log-Linear Regression Model

That is, the transformative log linear model shows constant elasticity. We can compare the original linear model and the transformed log linear model using the two slope coefficients, In linear model the slope of efficient β_2 gives the effect of a unit to change in X on the constant absolute change in Y. In log linear model the coefficient β_2 obtained from the model gives the constant percentage change in Y as a result of a 1% change in X.

We can compare the two models to compute an appropriate measure of the price elasticity. The price elasticity (E) is given as,

$$\beta_2 = (\frac{dY}{Y})/(\frac{dX}{X})$$

From this,

Slope = β_2 of the linear model.

In order to obtain price elasticity we have to multiply the slope with (X/Y). But the question is that which values of X and Y are taken? If we take the average values of X and Y ($\overline{x}$ and $\overline{y}$) for this purpose we have,

E = β_2 ($\overline{x}$ / $\overline{y}$)

But this result is something contrasts to the price elasticity derived from the log linear model. Therefore we always depends log linear model for calculating elasticity. The basic difference between a linear model and a log linear model is that for the log linear models slope and elasticity are the same but for a linear model, E = Slope (X/Y).

1.2.7.2 Semi-log Models

Semi-log models included log-lin model and lin-log models

a. Log-Lin Model and Measurement of Growth Rate

In economics, we are often interested in finding out the growth rates of GDP, population, money supply, employment, productivity, trade deficit etc. Log-lin models are very helpful in finding out these growth rates. We can explain this model as follows:

Suppose, we want to find out the rate of growth of real GDP over a period,

Let,

Y_t = real GDP at the time _t' Y_0 = Initial value of real GDP

By using the compound interest formula,

$$Y_t = Y_0(1+r)^t \quad \text{.....(1)}$$

Where, r = Compound rate of growth rate of Y

Taking natural logarithm on both sides,

$\ln Y_t = \ln Y_0 + t \ln (1+r)$..... (2)

Substituting,

$\beta_1 = \ln Y_0$ and $\beta_2 = \ln Y_t$

We have,

$\ln Y_t = \beta_1 + \beta_2 t$.....(3)

Adding the disturbance term we have, $\ln Y_t = \beta_1 + \beta_2 t + u_t$.....(4)

This model is also linear in parameters like any other regression models. The only difference is that the regressand is the logarithm of Y and the regressor is time which will take values 1, 2, 3 etc. That is, it is a semi log model because only one variable in the model (here Y the regressand) appears in the logarithmic form is called a log-lin model. One important property of this model is that the slope coefficient measures the constant proportional or relative change in Y ($\ln Y_t$: A change in the logarithm of a variable approximates the proportional (or percentage) change in that variable.)for a given absolute change in the values of the regressor, time.

That is, here,

$$\beta_2 = \frac{\text{Relative change in the Regressand}}{\text{Absolute change in the regressor}}$$

we multiply the relative change in Y by 100, it gives the percentage change or growth rate in Y for an absolute change in X.

A log-lin model like equation (4) is very useful where the X variable is time in some situations such as,

β_2 = Constant relative change in the variable y

$100(\beta_2)$ = Constant percentage change in the variable Y If $\beta_2 > 0$ = Rate of growth of variable Y

If $\beta_2 < 0$ = Rate of decay of the variable Y

That is why the models like equation (4) are called constant growth models. The growth rate obtained from log-lin models is the instantaneous rate of growth (rate of growth at a point of time). In order to calculate the compound growth rate,

Compound growth rate = $\{\text{Antilog}(\beta_2) - 1\}100$

This gives the compound growth rate over a period that we are considering for calculation.

b. Lin-log Model

A model in which the regressand (dependent variable) is linear but the regressors are logarithmic is called a lin log model. A lin-log model can be expressed as;

$Y_i = \beta_1 + \beta_2 \ln(X_i) + u_i (1)$

Lin-Log Model is used to find the absolute change in the dependent variable for a percentage in the independent variable whereas, the log-lin model used to find the percentage growth in the dependent variable for an absolute unit change in the independent variable.

In Lin-log model,

$Y_i = \beta_1 + \beta_2 \ln(X_i) + u_i$

We can interpret the slope coefficient β_2 as,

$$\beta_2 = \frac{\text{Change in Y}}{\text{Change in X}}$$

$$\beta_2 = \frac{\text{Change in Y}}{\text{Relative change in X}}$$

That is, a change in the log of a number is a relative change. Symbolically we have;

$$\beta_2 = \frac{\text{Absolute change in Y}}{\text{Absolute change in X}}$$

$$\frac{\Delta X}{\Delta X / X} \quad \text{--------- (2)}$$

That is,

$$\Delta Y = \beta_2\left(\frac{\Delta X}{X}\right) \text{ --------- (3)}$$

The equation (3) states that,

The absolute change in Y(ΔY) =β2 (relative change in X)

If the later term of the equation (3) is multiplied by 100 we have the absolute change in Y for a percentage change in X.

c. Reciprocal Models

Reciprocal Models of regression model are of the following type.

$$Y_t = B_1 + B_2\left(\frac{1}{X_1}\right) + u_i$$

That is, the dependent variable Y_t is a function of the reciprocal of the independent variable X_t. This model is non- linear in variable X because it enters inversely or reciprocally, the model is linear in β_1 and β_2 and is therefore a linear regression model.

The basic feature of a reciprocal model is that as X increases indefinitely the term β2 (1/X) approaches zero and Y approaches the limiting or asymptotic value of β1. The reciprocal models have built in them an asymptote or limit value that the dependent variable will take when the value of the X variable increases indefinitely. One of the important applications of the reciprocal model is Phillips curve. Some likely shapes of the curve corresponding to the reciprocal models are shown as in figure given below .

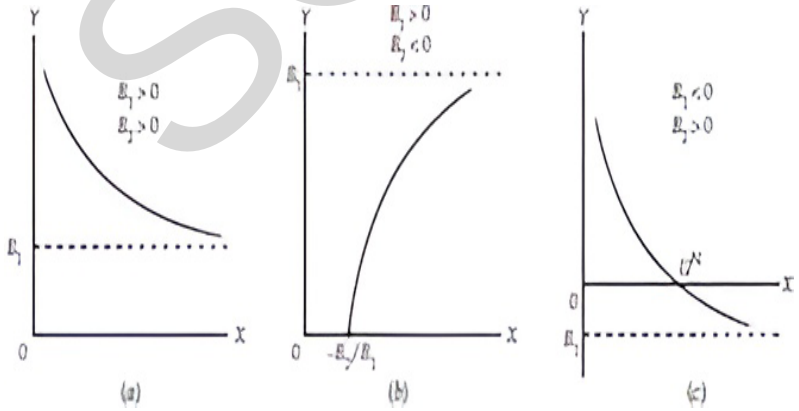


Fig.1.2.4. Reciprocal Models

Summarised Overview

The population regression function (PRF) represents the conditional mean of a dependent variable for given values of an explanatory variable. It shows how the average response of one variable changes as another variable changes. Using the example of income and consumption expenditure, the PRF is visualised as the population regression line, which indicates that average consumption rises with income. However, actual values often deviate from this mean, so the stochastic specification includes an error term, capturing random influences not explained by the model.

Since it is not always possible to study an entire population, the sample regression function (SRF) is used to estimate the PRF from sample data. The SRF provides estimates of the true population parameters and becomes closer to the PRF as the sample size increases. Estimation is typically done using the Ordinary Least Squares method, preferred for its simplicity, or the Maximum Likelihood method.

The error term in regression analysis has significant importance. It accounts for omitted or unknown variables, unobservable influences, randomness in human behaviour, measurement errors, and functional form misspecifications. It also allows the model to remain simple without losing too much explanatory power.

Regression analysis is broadly concerned with estimating the expected value of a dependent variable based on explanatory variables. A simple linear regression model involves one dependent and one independent variable, with the relationship assumed linear in parameters. This distinction between "simple" and "linear" ensures that the model remains mathematically tractable while capturing meaningful relationships.

In some cases, regression is specified through the origin, where the intercept is zero. An example is the capital asset pricing model (CAPM), in which risk premiums are expressed without an intercept term.

Beyond straight-line forms, different functional forms of regression models are used to capture specific economic relationships. Log–log or log-linear models measure constant elasticities, semi-log models (log–lin and lin–log) are useful for calculating growth rates and proportional changes, and reciprocal models incorporate asymptotic relationships, as in the Phillips curve. These transformations make complex relationships easier to estimate while allowing precise interpretation of elasticities, growth, and asymptotic behaviour.

Assignments

1. Define the Population Regression Function (PRF) and the Sample Regression Function (SRF). Why does the SRF vary from sample to sample?
2. Write a brief note on Simple Linear Regression Model
3. Under what economic circumstances is it reasonable to impose $\beta_0 = 0$? Give two concrete examples.
4. Explain elasticity interpretation under log linear model
5. Write the estimating equation for each functional form under log-lin model and lin-log model.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGrawHill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*): Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 3

OLS Estimation and Hypothesis Testing

Learning Outcomes

After completing this unit, the learner will be able to:

- understand estimation via Ordinary Least Squares (OLS)
- understand and evaluate Classical Assumptions or Gauss–Markov Theorem
- assess the concept of goodness of fit by incorporating R^2 and adjusted R^2.
- examine hypothesis testing under OLS framework
- equip with standard errors and deriving inference

Background

Preliminary idea regarding OLS as a statistical technique to estimate the parameters (coefficients) in a linear regression model is necessary so that it helps to conduct a detailed examination that under the OLS assumptions, the OLS estimator is the Best Linear Unbiased Estimator (BLUE). A close understanding of the observed and predicted values of the dependent variable also contributes to derive how well the regression model explains the variability of the dependent variable.

Keywords

Ordinary Least Squares (OLS), Best Fit Line, Unbiasedness, Variance, , Homoscedasticity, Correlation, Exogeneity, Normality, BLUE, Minimum Variance, R-squared, Sum of Squares (ESS Goodness of Fit, Hypothesis Testing

Discussion

Economists often find themselves standing at the intersection of theory and evidence, where they must translate abstract relationships into concrete, empirical statements. Ordinary Least Squares (OLS) is the primary vehicle for that translation. Beginning with a simple linear model, OLS answers the practical question: "Given the data at hand, which straight line best summarizes the average relationship between a dependent variable and one or more explanatory factors?" By minimizing the sum of squared residuals, OLS produces coefficient estimates that are easy to compute, intuitive to interpret, and—under a set of classical assumptions—statistically well-behaved.

1.3.1 Estimation through Ordinary Least Squares (OLS) Approach

The method of Ordinary Least Squares (OLS) was developed by a famous German mathematician Carl Frederich Gauss. The OLS method has some very attractive statistical properties under certain assumptions which made it one of the most powerful and popular method of regression analysis. The OLS method can be explained as follows.

Using OLS method through regression analysis we are estimating the PRF:

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

From SRF:

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

We have to know that;

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

$Yi = \hat{Y}i + \hat{u}i$ .....(5)

Where;

$\hat{Y}_i$ = estimated value or conditional mean value of Y_i

From equation (5) we have;

$\hat{u}_i = Y_i - \hat{Y}_i$ ..... (6) or

$\hat{u}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$ .....(7)

$\hat{u}_i = Y_i - (\hat{\beta}_0 - \hat{\beta}_1 X_i)$ .....(8)

It shows that the residuals ($\hat{u}_i$) are simply the differences between the actual and estimated Y values. For given _n' pairs of observations on Y and X, we would like to determine the SRF in such a manner that it is as close as possible to the actual Y. To attain this, we may use a criterion that the sum of residuals is as small as possible. That is,

$\sum \hat{u}_i = \sum (Y_i - \hat{Y}_i)$ is as small as possible.

But it is not a very good criterion because even though the residuals are not scattered evenly from the SRF we gave equal importance to each residuals. To clear this, we have to consider the following diagram (figure given below).

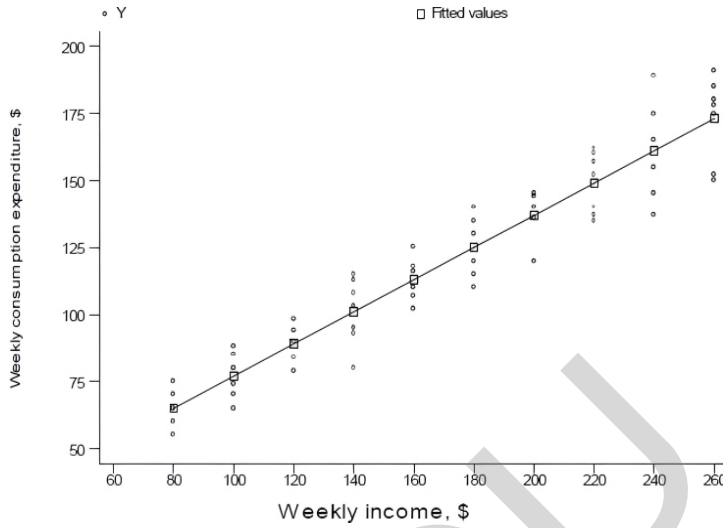


Fig.1.3.1. Scatter diagram of SRF and U_i

When we are considering the scatter diagram (figure given above) some residuals are closer to SRF whereas some others are widely scattered from the SRF. When we are adopting the criterion of minimizing $\sum \hat{u}_i$ by summing all the $\hat{u}_i$s, so that the algebraic sum of the $\hat{u}_i$ is small or even zero although the $\hat{u}_i$ are widely scattered about the SRF. Even if $\sum \hat{u}_i$ is small, we can find a greater difference between actual and estimated Y values. As a result, minimizing the sum of squares of errors is not considered to be a very good criterion to estimate PRF using SRF.

We can avoid this problem if we adopt the Least Square criterion. Using this criterion, the SRF can be fixed as;

$\sum \hat{u}_i^2 = \sum (Y_i - \hat{Y}_i)^2$ is as small as possible or

$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$ is as small as possible.

By squaring the residuals this method gives higher weightage to those residuals which are widely scattered about the SRF. So, in the OLS method the criteria adopted for fixing SRF is that the sum of squares of the residuals should be minimum to get best estimators.

The OLS principal chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ in such a manner that for a given sample $\sum \hat{u}i^2$ is as small as possible. In other words, for a given sample, the method of least squares provides us

with unique estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ that give the smallest possible value of $\sum \hat{u}i^2$.

Using OLS method the sum of squared residuals ($\sum \hat{u}i^2$) is to be minimised with respect to parameters. For this we are adopting two principles or two steps.

1. Differentiation principle and

2. Minimization principle

That is, we are first differentiating the residuals with respect to parameters. That is we are finding,

$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum \hat{u}_i^2\right)$ and $\frac{\partial}{\partial \hat{\beta}_1}\left(\sum \hat{u}_i^2\right)$

and then applying the minimization principle;

- First derivative should be equal to zero and

-Second derivative should be greater than zero or positive. That is the minimization principles are;

$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum \hat{u}_i^2\right) = 0$ and $\frac{\partial}{\partial \hat{\beta}_1}\left(\sum \hat{u}_i^2\right) = 0$

$\frac{\partial^2}{\partial \hat{\beta}_0^2}\left(\sum \hat{u}_i^2\right) > 0$ and $\frac{\partial^2}{\partial \hat{\beta}_1^2}\left(\sum \hat{u}_i^2\right) > 0$

By applying these two principles we arrive at two equations, popularly known as normal equations. The derivation of normal equation using differentiation and minimization principles are given below.

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

$$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

$$\hat{u}_i^2 = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Sum of squares,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} \hat{u}_i^2\right) = -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} \hat{u}_i^2\right) = 0 \Rightarrow -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

That is,

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} (X_i)$$

$$\frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} \hat{u}_i^2\right) = -2\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i$$

$$\frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} \hat{u}_i^2\right) = 0 \Rightarrow \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0$$

$$\sum Y_i X_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i) X_i = 0$$

That is,

$\sum Y_i X_i = \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X^2_i$ (2)

Then the two normal equations are;

$\Sigma Y_1 = n\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_1$ ---------- (1)

$\Sigma X_1 Y_1 = \hat{\beta}_0 \Sigma X_1 + \hat{\beta}_1 \Sigma X_1$ ---------- (2)

By solving these two normal equations we have

$$\hat{\beta}_1 = \frac{n\Sigma X\, i\, Y\, i - \Sigma X\, i \Sigma Y\, i}{n\Sigma X\, i^2 - (\Sigma X\, i)^2}$$

$$= \frac{(\Sigma X_i - \overline{X})(F_i - Y)}{(\Sigma X_i - \overline{X})^2}$$

$$= \frac{\Sigma x_1 y_1}{\Sigma x_i^2}$$

where, $x_i = X_i - \overline{X}$ and $y_i = Y_i - Y$

And,

$$\hat{\beta}_o = \frac{\sum Xi^2 \sum Yi - \sum Xi \sum YiXi}{n\sum Xi^2 - (\sum Xi)^2}$$

That is,

$$\hat{\beta}_o = \Upsilon - \hat{\beta}_1 \overline{X}$$

Therefore we have the two OLS estimators of the Simple Linear Regression model as,

$$\hat{\beta}_o = \Upsilon - \hat{\beta}_1 \overline{X} \text{ and}$$

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

$\hat{\beta}_1 \hat{\beta}_o$The estimators thus obtained (and) are known as the least square estimators or OLS estimators. This OLS estimators possess some statistical properties as;

-The OLS estimators are expressed solely in terms of observable quantities (sample quantities). Therefore, they can be easily computed.

-OLS estimators point estimators. That is, given the sample, each estimator will provide only single value of the relevant population parameter

-Once the OLS estimates are obtained from the sample data, the sample regression line can be easily obtained. The regression line thus obtained possess the following properties

- The regression line passes through the sample means of Y and X (X and Υ)
- The mean value of the estimated Y= $\hat{Y}_i$ which is equal to the mean value of the actual Y.
- The mean value of the residuals is zero
- The residuals $\hat{u}_i$ are uncorrelated with the predicted Y_i
- The residuals are uncorrelated with X.

1.3.2 Classical Linear Regression Model or Assumptions underlying the Method of OLS

Like many statistical analyses, ordinary least squares (OLS) regression has its own underlying assumptions. When these assumptions for linear regression are found true, Ordinary Least Squares produces the best estimates. However, if some of these assumptions are not materialised, you might need to employ remedial measures or use other estimation methods to improve the results.

Most of these assumptions pivot around the properties of the error term. Unfortunately, the error term is a population parameter that we never known in advance. Instead, we are using the next best thing that is available— the residuals. Residuals are the sample estimate of the error for each observation.

Assumption 1: The regression model is linear in parameters

This assumption addresses the functional form of the model. In statistics, linearity of a model can be expressed in two ways;

- Linearity in variables and
- Linearity in parameters

Linearity in variables means that the conditional expectation of Y is a linear function of Xi. That is, geometrically the regression curve in this case is a straight line. In short, the powers of the variables are always one. That is,

$E(Y/Xi) = \beta_0+\beta_1X_i$ is a linear function whereas, $E(Y/Xi) = \beta_0+\beta_1X_i^2$ is not a linear function.

Linearity in parameters means that the conditional expectation of Y is a linear function of parameters, the βs; it may or may not be linear in variables. That is the powers of βs are always one. For example, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$ is a regression model which are linear in parameters whereas, $Y=\beta_0+\beta_1^2X_i$ is not.

Of the two interpretations of linearity, linearity in parameters is relevant for the development of regression theory.

In fact, the defining characteristic of linear regression is this functional form of the *parameters* rather than the ability to model curvature. Linear models can model curvature by including nonlinear *variables* such as polynomials and transforming exponential functions. To satisfy this assumption, the correctly specified model must fit the linear pattern.

Assumption 2: The error term has a population mean of zero

The error term in fact explains the variation in the dependent variable that the independent variables do not explain. Random chance should determine the values of the error term. For our model to be unbiased, the average value of the error term must equal zero. That is, $E(U_i/X_i) = 0$.

Suppose the average error is +7. This non-zero average error indicates that our model systematically under-estimates the observed values. Statisticians refer to systematic error like this as bias, and it signifies that our model is inadequate because it is not correct on an average.

Assumption 3: Observations of the error term are unorrelated with each other (No Autocorrelation)

This assumption says that no two observations of the error term in a regression model are correlated. That is, one observation of the error term should not predict the next observation. For instance, if the error for one observation is positive and that systematically increases the probability that the following error is positive, that is a positive autocorrelation. If the

subsequent error is more likely to have the opposite sign, that is a negative autocorrelation. This problem is known both as serial correlation and autocorrelation. Serial correlation is most likely to occur in time series models. Symbolically no autocorrelation can be expressed as,

$Cov(U_i,U_j / X_i,X_j) = 0$

Assess this assumption by graphing the residuals in the order that the data were collected. We want to see randomness in the plot. In the graph for a sales model, there is a cyclical pattern with a positive autocorrelation.

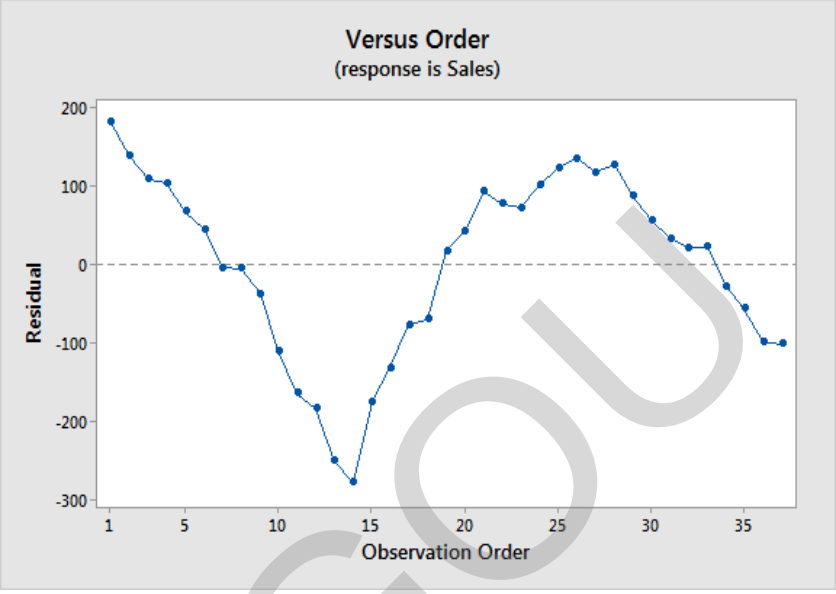


Fig.1.3.2. Autocorrelation

If we have information that allows to predict the error term for an observation, we must incorporate that information into the model itself. To resolve this issue, we might need to add an independent variable to the model that captures this information. For the sales model shown in figure given above, we need to add variables that explain the cyclical pattern.

Serial correlation reduces the precision of OLS estimates. Analysts can also use time series analysis for time dependent effects.

Assumption 4: The error term has a constant variance (No Hetroscedasticity)

The error variance should be consistent for all observations. In other words, the variance does not change for each observation or for a range of observations. This preferred condition is known as homoscedasticity (same spread/scatter). If the variance changes, we refer to that as heteroscedasticity (different spread/scatter). Homoscedasticity can be expressed symbolically as,

$Var(U_i/X_i) = \sigma^2$

Whereas, the hetroscedasticity can be expressed as, $Var(U_i/X_i) = \sigma_i^2$

The easiest way to check this assumption is to create residuals versus estimated value plot. On

this type of graph, heteroscedasticity appears as a cone shape where the spread of the residuals increases in one direction. In the graph below (figure given below), the spread of the residuals increases as the fitted value increases.

Heteroscedasticity reduces the precision of the estimates in OLS linear regression.

When both assumption 4 (no autocorrelation) and 5 (homoscedasticity) are true, statisticians say that the error term is independent and identically distributed (IID).

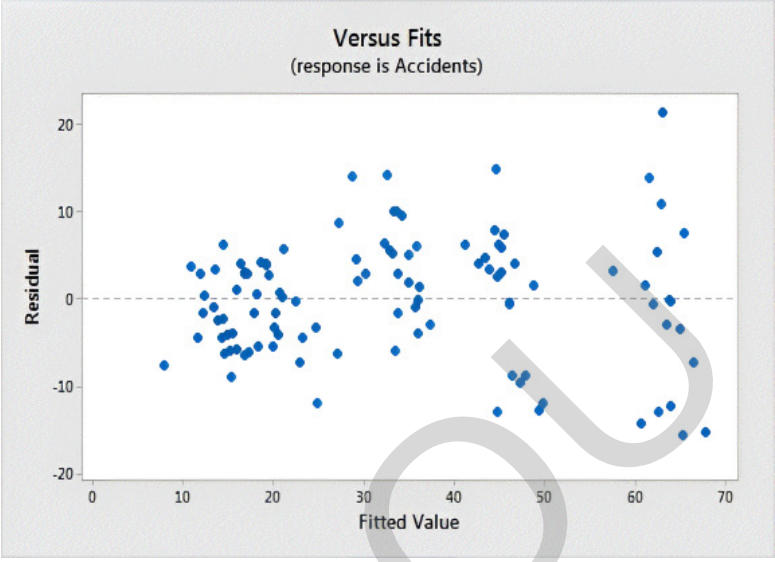


Fig.1.3.3. Hetroscedasticity

Assumption 5: No independent variable is a perfect linear function of other explanatory variables (No perfect Multicollinearity)

Perfect correlation occurs when two variables have a Pearson's correlation coefficient of +1 or -1. Under such circumstances, when one of the variable changes, the other variable also changes in a fixed proportion.

Ordinary Least Squares cannot distinguish one variable from the other when they are perfectly correlated. If these correlations are very high, it can cause problems. Statisticians refer to this condition as multicollinearity, and it reduces the precision of the estimation in OLS linear regression.

Assumption 6: The error term is normally distributed

For estimation of the regression model using OLS does not require the assumption that the error term follows a normal distribution to produce unbiased estimates with the minimum variance. However, satisfying this assumption allows us to perform statistical hypothesis testing and generate reliable confidence intervals and prediction intervals.

The easiest way to determine whether the residuals follow a normal distribution is to assess a normal probability plot. If the residuals follow the straight line on this type of graph, they are normally distributed (figure given below)

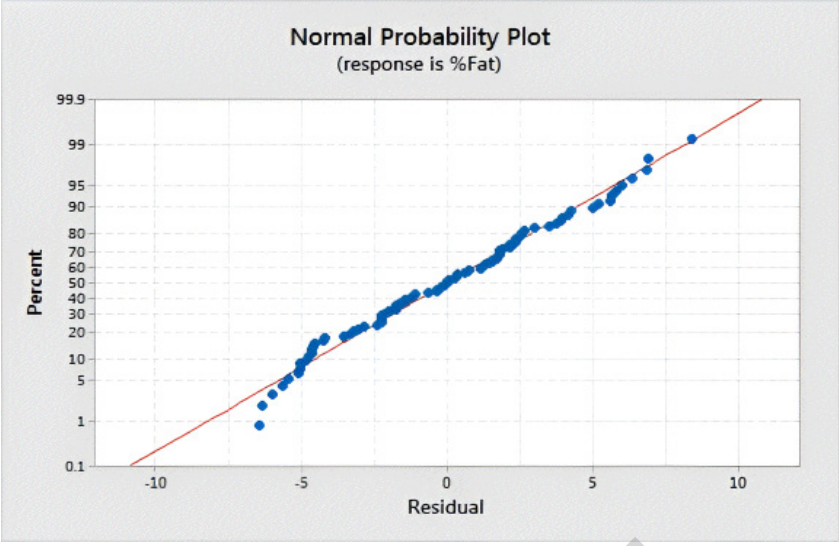


Fig.1.3.4. Normal probability plot

Assumption 7: X values are fixed in repeated sampling

Values taken by the Independent variable X (regressor) are considered to be fixed in repeated sampling. More technically X is assumed to be non-stochastic. When we are estimating PRF based on SRF we may relies on more than one sample. In all these repeated sampling the values of the regressor will change. This means that our regression analysis is conditional regression analysis, ie, conditional on the given values of the regressor X. That is we are estimating $E(Y/X_i)$.

Assumption 8: Zero covariance between U_i and X_i

This assumption states that the explanatory variables X and the disturbance U are uncorrelated. It is because in our PRF we are assuming that X and U have separate influence on Y. When X and U are correlated it is not possible to assess their individual effects on Y. This zero covariance between X and U can be expressed as;

$Cov(U_i,X_i) = 0$

Assumption 9: The number of observations 'n' must be greater than the number of parameters to be estimated

This assumption states that the number of observations associated with X and Y should be greater than the number of parameters (βs) to be estimated. Alternately, the number of observations 'n' must be greater than the number of explanatory variable Xs.

Assumption 10: Variability in X values

This assumption is very important. This assumption states that

$\overline{XX}$X values are not identical. That is $X_i \neq \overline{X}$. When $X_i = \overline{X}$, we cannot measure the β_2 and hence the variability in Y. In short, the X values in a given sample must not all be the same. Technically variance of X must be a finite positive number.

Var (X)>0

Simply the assumption states that there must be variability both in X and Y values.

Assumption 11: Regression model is correctly specified

The CLRM assumed that the model used to test an economic theory is correctly specified. Alternately there is no specification bias or errors in the model used in regression analysis. An econometric investigation begins with the specification of the model underlying the phenomenon of interest. The model specification includes,

-The selection of variables to be included

-The selection of the functional form of the model

-What are the assumptions made about X_i, Y_i and U_i

By doing all these correctly we have valid estimates. The validity of interpreting the estimated regression is highly questionable when the used models are of wrong functional form. Therefore, the correct specification of the economic model is of great importance.

1.3.3 Properties of OLS Estimators Or The Gauss-Markov Theorem

Linear regression models have several applications in real life. In econometrics, Ordinary Least Squares (OLS) method is widely used to estimate the parameters of a linear regression model. For the validity of OLS estimates, the following assumptions are made while running linear regression models.

A1. The linear regression model is —linear in parameters.

A2. There is a random sampling of observations.

A3. The conditional mean should be zero.

A4. There is no multi-collinearity (or perfect collinearity)

A5. Spherical errors: There is homoscedasticity and no autocorrelation

A6: Optional Assumption: Error terms should be normally distributed.

These assumptions are extremely important because violation of any of any of these assumptions would make OLS estimates unreliable and incorrect. Specifically, a violation would result in incorrect signs of OLS estimates, or the variance of OLS estimates would be unreliable, leading to confidence intervals that are too wide or too narrow.

Hence, it is necessary to investigate why OLS estimators and its assumptions gather so much focus. This can be answered by evaluating the properties of OLS model and the famous Gauss-Markov Theorem.

1.3.3.1 Gauss-Markov Theorem

The Gauss-Markov Theorem is named after Carl Friedrich Gauss and Andrey Markov.

Let the regression model be:

$Yi = \beta_0+\beta_1X_i + u_i$

Let $\hat{\beta_0}$ is the estimator of β_0 and $\hat{\beta_1}$ is the estimator of β_1. According to the Gauss-Markov Theorem, under the assumptions A_1 to A_5 of the linear regression model, the OLS estimators $\hat{\beta_0}$ and $\hat{\beta_1}$ are the Best Linear Unbiased Estimators (BLUE) of β_0 and β_1. In other words, the OLS estimators $\hat{\beta_0}$and $\hat{\beta_1}$have the minimum variance of all linear and unbiased estimators of β_0 and β_1. BLUE summarizes the properties of OLS regression. These properties of OLS in econometrics are extremely important, thus making OLS estimators one of the strongest and most widely used estimators for unknown parameters. This theorem tells that one should use OLS estimators not only because it is unbiased but also because it has minimum variance among the class of all linear and unbiased estimators.

1.3.3.2 Properties of OLS Regression Estimators

Property 1: Linear

This property is more concerned with the estimator rather than the original equation that is being estimated. In assumption A_1, the focus was that the linear regression should be —linear in parameters. However, the *linear* property of OLS estimator means that OLS belongs to that class of estimators, which are linear in Y, the dependent variable. Note that OLS estimators are linear only with respect to the dependent variable and not necessarily with respect to the independent variables. The *linear* property of OLS estimators doesn't depend only on assumption A_1 but on all assumptions A_1 to A_5.

Proof:-

We have,

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

$$= \frac{\sum_i x_i (Y_i - \bar{Y})}{\sum_i x_i^2}$$

$$= \frac{\sum_i x_i Y_i}{\sum_i x_i^2} - \frac{\bar{Y}\sum_i x_i}{\sum_i x_i^2}$$

$$= \frac{\sum_i x_i Y_i}{\sum_i x_i^2} \qquad \text{because } \sum_i x_i = 0.$$

Defining the observation weights $\mathbf{k_i = x_i / \sum_i x_i^2}$ for i = 1, …, N, we can rewrite the last expression above for as:

$$\hat{\beta}_1 = \sum_i k_i Y_i \qquad \text{where } k_i \equiv \frac{x_i}{\sum_i x_i^2} \qquad (i = 1, ..., N)$$

Note that the formula for $\hat{\beta}_1$ and the definition of the weights k_i imply that is also a linear function of the Y_i such that

$$\hat{\beta}_1 = \sum_i k_i y_i$$

The OLS slope coefficient estimator $\hat{\beta}_1$ is a linear function of the sample values Y_i (i = 1,…,N), where the coefficient of Yi or yi is ki.

Properties of the weights k_i

In order to establish the remaining properties of $\hat{\beta}_1$, it is necessary to know the arithmetic properties of the weights ki. These properties are,

[K1] $\mathbf{\sum_i k_i = 0}$, i.e., the weights k_i sum to zero.

$$\sum_i k_i = \sum_i \frac{x_i}{\sum_i x_i^2} = \frac{1}{\sum_i x_i^2} \sum_i x_i = 0, \qquad \text{because } \sum_i x_i = 0.$$

[K2] $\mathbf{\sum_i k_i^2 = \frac{1}{\sum_i x_i^2}}$.

$$\sum_i k_i^2 = \sum_i \left(\frac{x_i}{\sum_i x_i^2} \right)^2 = \sum_i \frac{x_i^2}{\left(\sum_i x_i^2\right)^2} = \frac{\left(\sum_i x_i^2\right)}{\left(\sum_i x_i^2\right)^2} = \frac{1}{\sum_i x_i^2}.$$

[K3] $\mathbf{\sum_i k_i x_i = \sum_i k_i X_i}$.

$$\begin{aligned} \sum_i k_i x_i &= \sum_i k_i (X_i - \bar{X}) \\ &= \sum_i k_i X_i - \bar{X} \sum_i k_i \\ &= \sum_i k_i X_i \qquad \text{since } \sum_i k_i = 0 \text{ by [K1] above.} \end{aligned}$$

[K4] $\mathbf{\sum_i k_i x_i = 1}$.

$$\sum_i k_i x_i = \sum_i \left(\frac{x_i}{\sum_i x_i^2} \right) x_i = \sum_i \frac{x_i^2}{\left(\sum_i x_i^2\right)} = \frac{\left(\sum_i x_i^2\right)}{\left(\sum_i x_i^2\right)} = 1.$$

Implication: $\mathbf{\sum_i k_i X_i = 1.}$

Property 2: Unbiasedness

Looking at the regression equation, you will find an error term associated with the regression equation that is estimated. This makes the dependent variable also random. If an estimator uses the dependent variable, then that estimator would also be a random number. Therefore, before describing what unbiasedness is, it is important to mention that unbiasedness property is a property of the estimator and not of any sample.

Unbiasedness is one of the most desirable properties of any estimator. The estimator should ideally be an unbiased estimator of true parameter/population values.

Consider a simple example: Suppose there is a population of size 1000, and you are taking out samples of 50 from this population to estimate the population parameters. Every time you take a sample, it will have the different set of 50 observations and, hence, you would estimate different values of of $\hat{\beta}_0$ and $\hat{\beta}_1$

The unbiasedness property of OLS method says that when you take out samples of 50 repeatedly, then after some repeated attempts, you would find that the average of all the $\hat{\beta}_0$ and $\hat{\beta}_1$ from the samples will equal to the actual (or the population) values of β0 and β1.

The OLS coefficient estimator $\hat{\beta}_1$ is unblased, meaning that $E(\hat{\beta}_1) = \hat{\beta}_1$

The OLS coefficient estimator $\hat{\beta}_0$ is unblased, meaning that $E(\hat{\beta}_0) = \hat{\beta}_0$

Here, 'E' is the expectation operator.

If you take out several samples, keep recording the values of the estimates, and then take an average, you will get very close to the correct population value. If your estimator is biased, then the average will not equal the true parameter value in the population.

Proof:-

We have,

$$\hat{\beta}_1 = \sum_i k_i Y_i$$

Our PRF is, $Y_i = \beta_0 + \beta_1 X_i + u_i$

$$\begin{aligned}
\hat{\beta}_1 &= \sum_i k_i Y_i \\
&= \sum_i k_i(\beta_0 + \beta_1 X_i + u_i) && \text{since } Y_i = \beta_0 + \beta_1 X_i + u_i \text{ by A1} \\
&= \beta_0 \sum_i k_i + \beta_1 \sum_i k_i X_i + \sum_i k_i u_i \\
&= \beta_1 + \sum_i k_i u_i, && \text{since } \sum_i k_i = 0 \text{ and } \sum_i k_i X_i = 1.
\end{aligned}$$

Now take expectations of the above expression for $\hat{\beta}_1$, conditional on the sample values {Xi: i = 1, …, N} of the regressor X. Conditioning on the sample values of the regressor X means that the ki are treated as nonrandom, since the ki are functions only of the Xi.

$$\begin{aligned}
E(\hat{\beta}_1) &= E(\beta_1) + E[\sum_i k_i u_i] \\
&= \beta_1 + \sum_i k_i E(u_i | X_i) && \text{since } \beta_1 \text{ is a constant and the } k_i \text{ are nonrandom} \\
&= \beta_1 + \sum_i k_i 0 && \text{since } E(u_i | X_i) = 0 \text{ by assumption A2} \\
&= \beta_1.
\end{aligned}$$

Result: The OLS slope coefficient estimator $\hat{\beta}_1$ is an unbiased estimator of the slope coefficient β_1: that is,

$E(\hat{\beta}_1) = \hat{\beta}_1$

Graphically the property of unbiasedness is depicted in figure given below. The unbiasedness property of OLS in Econometrics is the basic minimum requirement to be satisfied by any

estimator. However, it is not sufficient for the reason that most times in real-life applications, we will not have the luxury of taking out repeated samples. In fact, only one sample will be available in most cases.

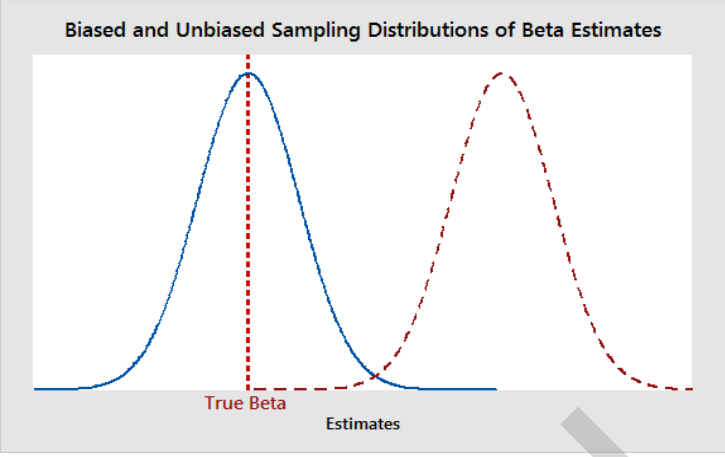


Fig. 1.3.5. Unbiasedness of OLS Estimators

Property 3: Best: Minimum Variance

The efficient property of any estimator says that the estimator is the minimum variance unbiased estimator. Therefore, if we take all the unbiased estimators of the unknown population parameter, the estimator will have the least variance. The estimator that has less variance will have individual data points closer to the mean. As a result, they will be more likely to give better and accurate results than other estimators having higher variance. Thus:

1. If the estimator is unbiased but doesn't have the least variance – it is not the best.
2. If the estimator has the least variance but is biased –it is again not the best.
3. If the estimator is both unbiased and has the least variance – it is the best estimator.

OLS estimators have the least variance among the class of all linear unbiased estimators. So, this property of OLS regression is less strict than efficiency property. Efficiency property says least variance among all unbiased estimators, and OLS estimators have the least variance among all linear and unbiased estimators.

Proof:-

The variance of the OLS slope coefficient estimator $\hat{\beta}_1$ is defined as;

Since $\hat{\beta}_1$ is an unbiased estimation of $\hat{\beta}_1$, E ($\hat{\beta}_1$) = $\hat{\beta}_1$. The variance of $\hat{\beta}_1$ can therefore be written as

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

From part (1) of the unbiasedness proofs above, the term $[\hat{\beta}_1 - \beta_1]$, which is called the **sampling error of $\hat{\beta}_1$**, is given by

$$\mathrm{Var}(\hat{\beta}_1) \equiv E\left\{\left[\hat{\beta}_1 - E(\hat{\beta}_1)\right]^2\right\}. \qquad \left[\hat{\beta}_1 - \beta_1\right] = \Sigma_i k_i u_i .$$

The square of the sampling error is therefore

$$\left[\hat{\beta}_1 - \beta_1\right]^2 = (\Sigma_i k_i u_i)^2$$

Since the square of a sum is equal to the sum of the squares plus twice the sum of the cross products,

$$\left[\hat{\beta}_1 - \beta_1\right]^2 = (\Sigma_i k_i u_i)^2 = \sum_{i=1}^{N} k_i^2 u_i^2 + 2\sum_{i<s}\sum_{s=2}^{N} k_i k_s u_i u_s .$$

Now use assumptions A3 and A4 of the classical linear regression model (CLRM):

(A3) $\mathrm{Var}(u_i | X_i) = E(u_i^2 | X_i) = \sigma^2 > 0$ for all i = 1, ..., N;

(A4) $\mathrm{Cov}(u_i, u_s | X_i, X_s) = E(u_i u_s | X_i, X_s) = 0$ for all $i \neq s$.

We take expectations conditional on the sample values of the regressor X:

$$E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] = \sum_{i=1}^{N} k_i^2 E(u_i^2 | X_i) + 2\sum_{i<s}\sum_{s=2}^{N} k_i k_s E(u_i u_s | X_i, X_s)$$

$$= \sum_{i=1}^{N} k_i^2 E(u_i^2 | X_i) \quad \text{since } E(u_i u_s | X_i, X_s) = 0 \quad \text{for } i \neq s \text{ by (A4)}$$

$$= \sum_{i=1}^{N} k_i^2 \sigma^2 \quad \text{since } E(u_i^2 | X_i) = \sigma^2 \ \forall i \text{ by (A3)}$$

$$= \frac{\sigma^2}{\Sigma_i x_i^2} \quad \text{since } \Sigma_i k_i^2 = \frac{1}{\Sigma_i x_i^2} \text{ by (K2).}$$

Result: The ***variance*** of the OLS slope coefficient estimator $\hat{\beta}_1$ is

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma_i x_i^2} = \frac{\sigma^2}{\Sigma_i (X_i - \bar{X})^2} = \frac{\sigma^2}{\mathrm{TSS}_X} \quad \text{where } \mathrm{TSS}_X = \Sigma_i x_i^2$$

The ***standard error*** of $\hat{\beta}_1$ is the square root of the variance: i.e.,

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\mathrm{Var}(\hat{\beta}_1)} = \left(\frac{\sigma^2}{\Sigma_i x_i^2}\right)^{\frac{1}{2}} = \frac{\sigma}{\sqrt{\Sigma_i x_i^2}} = \frac{\sigma}{\sqrt{\mathrm{TSS}_X}}.$$

The minimum variance property of OLS estimators can be graphically plotted as shown in figure given below.

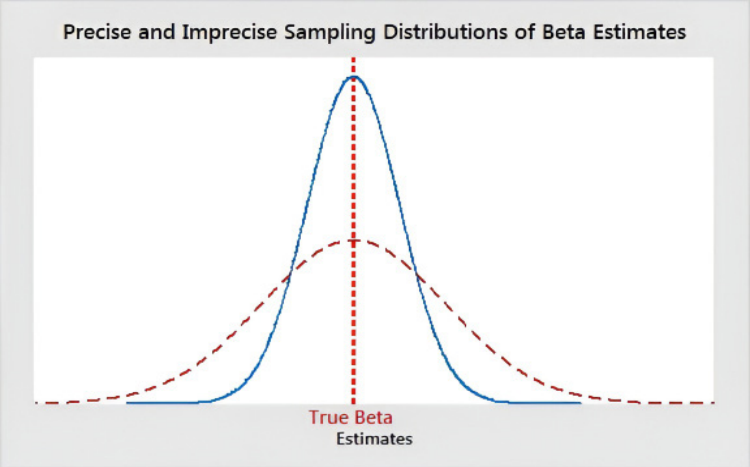


Fig.1.3.6. Minimum Variance of OLS Estimator

The above three properties of OLS model makes OLS estimators BLUE as mentioned in the Gauss-Markov theorem. It is worth spending time on some other estimators' properties of OLS in econometrics. The properties of OLS described below are asymptotic properties of OLS estimators. So far, finite sample properties of OLS regression were discussed. These properties tried to study the behaviour of the OLS estimator under the assumption that you can have several samples and, hence, several estimators of the same unknown population parameter. In short, the properties were that the average of these estimators in different samples should be equal to the true population parameter (unbiasedness), or the average distance to the true parameter value should be the least (efficient). However, in real life, you will often have just one sample. Hence, asymptotic properties of OLS model are discussed, which studies how OLS estimators behave as sample size increases. In this case, sample size should be large.

Property 4: Asymptotic Unbiasedness

This property of OLS says that as the sample size increases, the biasedness of OLS estimators disappears.

Property 5: Consistency

An estimator is said to be consistent if its value approaches the actual, true parameter (population) value as the sample size increases. An estimator is consistent if it satisfies two conditions:

a. It is asymptotically unbiased

b. Its variance converges to 0 as the sample size increases.

Both these hold true for OLS estimators and, hence, they are consistent estimators. For an estimator to be useful, consistency is the minimum basic requirement. Since there may be several such estimators, asymptotic efficiency also is considered. Asymptotic efficiency is the sufficient condition that makes OLS estimators the best estimators.

To conclude, linear regression is important and widely used, and OLS estimation technique is the most prevalent. OLS estimators are BLUE (i.e. they are linear, unbiased and efficient, have

the least variance among the class of all linear and unbiased estimators). Amidst all this, one should not forget the Gauss-Markov Theorem (i.e. the estimators of OLS model are BLUE) holds only if the assumptions of OLS are satisfied.

Each assumption that is made while studying OLS adds restrictions to the model, but at the same time, also allows to make stronger statements regarding OLS. So, whenever we are planning to use a linear regression model using OLS, always check for the OLS assumptions. If the OLS assumptions are satisfied, then life becomes simpler, for we can directly use OLS for the best results.

1.3.4 Coefficient of Determination/ Goodness of Fit (r2)

Here we are considering the goodness of fit of the fitted regression line to a set of data. That is, we shall find out how well the sample regression line fits the data. For this we are using the concept 'coefficient of determination'. The coefficient of determination, denoted by r^2, is the proportion of the variations in the dependent variable that is predictable from the independent variable. The coefficient of determination is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event. In other words, this coefficient, which is more commonly known as r^2, assesses how strong the linear relationship is between two variables. The coefficient of determination is used to explain how much variability of one factor can be caused by its relationship to another factor.

It is clear that, if all the observations were lie on the regression line, we would obtain a perfect fit. But it is a rare case. This coefficient is commonly known as r^2 and is sometimes referred to as the "goodness of fit." r^2 is simply the square of the sample correlation coefficient (i.e., r) between the observed outcomes and the observed predictor values. This measure is represented as a value between 0.0 and 1.0 ($0 \leq r2 \leq 1$), where a value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future forecasts, while a value of 0.0 would indicate that the model fails to accurately model the data at all.

We can show the goodness of fit of a regression line through the graph (figure given below) and from that we can calculate the value of r^2.

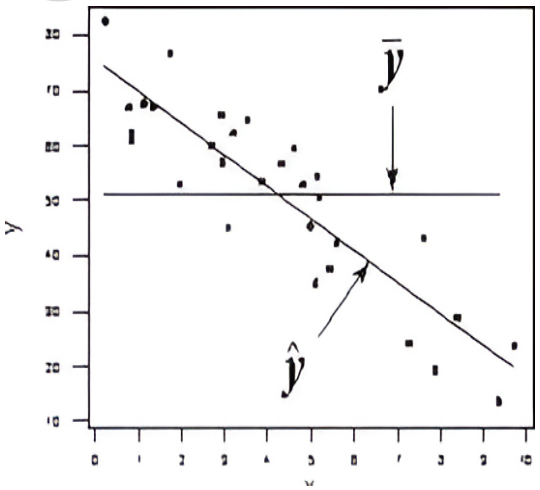


Fig.1.3.7. Goodness of Fit of Estimated Regression Line

There are two lines in figure a horizontal line placed at the average response, $\bar{y}$, and a shallow-sloped estimated regression line, $\hat{y}$. From figure the calculation of sum of squares are;

- ***Explained Sum of Squares (ESS)*** quantifies how far the estimated sloped regression line, , is from the horizontal "no relationship line," the sample mean or $\bar{y}$.

  That is, ESS= $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

- ***Residual sum of Squares (RSS)*** quantifies how much the data points, yi, vary around the estimated regression line, $\hat{y}$.

  That is, RSS = $\sum_{i=1}^{n}(y_i - \hat{y})^2$

- ***Total Sum of Squares (TSS)*** quantifies how much the data points, yi, vary around their mean, $\bar{y}$

  That is, TSS = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

The sum of squares can be better illustrated in figure given below. From these sum of squares, r^2 can be calculated as,

r^2 = ESS/TSS

That is, $r^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Or in other words,

r^2 = 1- (RSS/TSS) since ESS+RSS= TSS

Therefore; $\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \sum_{i=1}^{n}(y_i - \hat{y})^2$

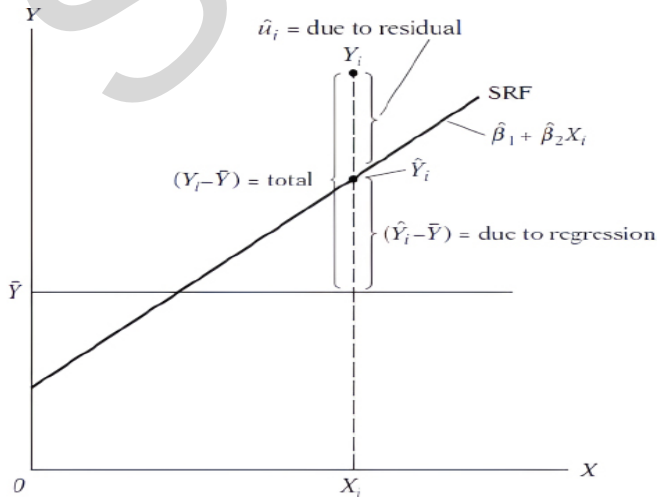


Fig. 1.3.8. Sum of Squares

Some basic characteristics of the measure are as follows:

- Since r^2 is a proportion, it is always a number between 0 and 1.
- If $r^2 = 1$, all of the data points fall perfectly on the regression line. The predictor x accounts for *all* of the variation in yi
- If $r^2 = 0$, the estimated regression line is perfectly horizontal. The predictor x accounts for *none* of the variation in yi

We've learned the interpretation for the two easy cases — when r2 = 0 or r2 = 1 — but, how do we interpret r2 when it is some number between 0 and 1. In this situation, the coefficient of determination r2 can be interpreted as, "r2 ×100 percent of the variation in Y is explained by the variation in predictor X."

1.3.5 Estimation and Hypothesis Testing

One important way to make statistical inferences about a population parameter, we use hypothesis testing to make decisions about the parameter's value. Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. Hypothesis is in fact an if -then proposition. The methodology employed for hypothesis testing depends on the nature of the data used and the objectives to be resolved. Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process.

The null hypothesis (null always indicates zero) is usually a hypothesis of equality between population parameters; e.g., a null hypothesis may state that the population mean is equal to

zero. The alternative hypothesis is effectively the opposite of a null hypothesis (e.g., the population mean return is not equal to zero). Thus, they are mutually exclusive, and only one can be true. However, one of the two hypotheses will always be true.

There are mainly two ways for proceeding with the testing of a hypothesis.

1. The Rejection Region Method

To decide between two competing claims, we can conduct a hypothesis test as follows.

- Express the claim about a specific value for the population parameter of interest as a null hypothesis, denoted H0. The null hypothesis needs to be in the form "parameter = some hypothesized value," for example, H0: E(Y) = 255.
- Express the alternative claim as an alternative hypothesis, denoted H1. The alternative hypothesis can be in a lower tail form, for example, H1: E(Y) < 255, or an upper tail form, for example, H1: E(Y) > 255, or a two-tail form, for example, H1: E(Y) ≠ 255. The alternative hypothesis, also sometimes called the research hypothesis, is what we would like to demonstrate to be the case, and needs to be stated before looking at the data.

- Calculate a test statistic based on the assumption that the null hypothesis is true. For testing a univariate population mean, the relevant test statistic is t-statistic.
- Under the assumption that the null hypothesis is true, this test statistic will have a particular probability distribution. For testing a univariate population mean, this t-statistic has a t-distribution with n−1 degrees of freedom. We would therefore expect it to be "close" to zero (if the null hypothesis is true). Conversely, if it is far from zero, then we might begin to doubt the null hypothesis:
    - For an upper-tail test, a t-statistic that is positive and far from zero would then lead us to favour the alternative hypothesis (a t-statistic that was far from zero but negative would favour neither hypothesis and the test would be inconclusive).
    - For a lower-tail test, a t-statistic that is negative and far from zero would then lead us to favour the alternative hypothesis (a t-statistic that was far from zero but positive would favour neither hypothesis and the test would be inconclusive).
    - For a two-tail test, any t-statistic that is far from zero (positive or negative) would lead us to favour the alternative hypothesis.
- There is always a chance that we might mistakenly reject a null hypothesis when it is actually true. Often, this chance—called the Level of significance- will be set at 5%, but more stringent tests (such as in clinical trials of new pharmaceutical drugs) might set this at 1%, while less stringent tests (such as in sociological studies) might set this at 10%. For the sake of argument, we use 5% as a default value for hypothesis tests in this course (unless stated otherwise).
- The significance level dictates the critical value(s) for the test, beyond which an observed t-statistic leads to rejection of the null hypothesis in favour of the alternative. This region, which leads to rejection of the null hypothesis, is called the rejection region. For example, for a significance level of 5%:
    - For an upper-tail test, the critical value is the 95th percentile of the t-distribution with n−1 degrees of freedom; reject the null in favour of the alternative if the t statistic is greater than this.
    - For a lower-tail test, the critical value is the 5th percentile of the t-distribution with n−1 degrees of freedom; reject the null in favour of the alternative if the t-statistic is less than this.
    - For a two-tail test, the two critical values are the 2.5th and the 97.5th percentiles of the t-distribution with n−1 degrees of freedom; reject the null in favour of the alternative if the t-statistic is less than the 2.5th percentile or greater than the 97.5th percentile.

2. The p-value Method

An alternative way to conduct a hypothesis test, firstly we assume again that the null hypothesis is true, but then to calculate the probability of observing a t-statistic as extreme as the one

observed or even more extreme (in the direction that favours the alternative hypothesis). This is known as the p value (sometimes also called the observed significance level):

- For an upper-tail test, the p-value is the area under the curve of the t-distribution (with n−1 degrees of freedom) to the right of the observed t-statistic.
- For a lower-tail test, the p-value is the area under the curve of the t-distribution (with n−1 degrees of freedom) to the left of the observed t-statistic.
- For a two-tail test, the p-value is the sum of the areas under the curve of the t-distribution (with n−1 degrees of freedom) beyond both the observed t-statistic and the negative of the observed t-statistic.

If the p-value is too "small," then this suggests that it seems unlikely that the null hypothesis could have been true—so we reject it in favour of the alternative. Otherwise, the t-statistic could well have arisen while the null hypothesis held true—so we do not reject it in favour of

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

the alternative. Again, the significance level chosen tells us how small is small: If the p-value is less than the significance level, then reject the null in favour of the alternative; otherwise, do not reject it.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3.'t' Test

The' t test is usually used to conduct hypothesis tests on the regression coefficients (βs) obtained from simple linear regression. A statistic based on the _t' distribution is used to test the two-sided hypothesis that the true slope, β_1, equals some constant value, $\beta_{1,0}$. The statements for the hypothesis test are expressed as:

$$H_0: \quad \beta_1 = \beta_{1,0}$$
$$H_1: \quad \beta_1 \neq \beta_{1,0}$$

The test statistic used for this test is:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

where $\hat{\beta}_1$ is the least square estimate of β1, and se($\hat{\beta}_1$) is its standard error. The value of se($\hat{\beta}_1$) can be calculated as follows:

$$se(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

The test statistic, T0 , follows a t distribution with (n−2) degrees of freedom, where n is the total number of observations. The null hypothesis, H0, is accepted if the calculated value of the test statistic is such that:

$$-t_{\alpha/2,n-2} < T_0 < t_{\alpha/2,n-2}$$

where tα/2,n−2 and −tα/2,n−2 are the critical values for the two-sided hypothesis. tα/2,n−2 is the percentile of the t distribution corresponding to a cumulative probability of (1−α/2) and α is the significance level.

If the value of $\beta_{1,0}$ is zero, then the hypothesis tests for the significance of regression. In other words, the test indicates if the fitted regression model is significant in explaining variations in the observations or if you are trying to impose a regression model when no true relationship exists between x and Y. Failure to reject H0:β1=0 implies that no linear relationship exists between x and Y. This result may be obtained when the scatter plots of against are as shown as figure given below.

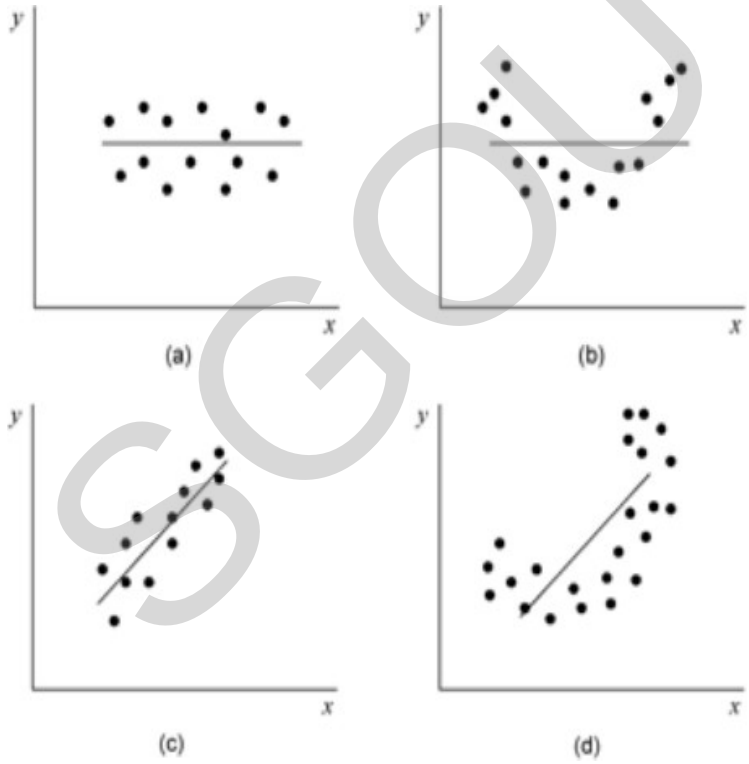


Fig. 1.3.9 Scatter Plots

In figure given above, figure (a) represents the case where no model exits for the observed data. In this case you would be trying to fit a regression model to noise or random variation. (b) represents the case where the true relationship between x and Y is not linear. (c) and (d) represent the case when H0:β1=0 is rejected, implying that a model does exist between x and Y. (c) represents the case where the linear model is sufficient. In the following figure, (d) represents the case where a higher order model may be needed.

A similar procedure can be used to test the hypothesis on the intercept. The test statistic used in this case is:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

where $\hat{\beta}_0$ is the least square estimate of β0, and se($\hat{\beta}_0$) is its standard error which is calculated using:

$$se(\hat{\beta}_0) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

4. F Test

***F*-test** is any statistical test in which the test statistic follows an *F*-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact "*F*- tests" mainly arise when the models have been fitted to the data using least squares. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s. Common examples of the use of *F*-tests include the study of the following cases:

- For checking the overall significance of the fitted regression model.
- The hypothesis that the means of a given set of normally distributed populations, all having the same standard deviation, are equal. This is perhaps the best-known *F*-test, and plays an important role in the analysis of variance (ANOVA).
- The hypothesis that a proposed regression model fits the data well. See Lack-of-fit sum of squares.
- The hypothesis that a data set in a regression analysis follows the simpler of two proposed linear models that are nested within each other.

In addition, some statistical procedures, such as Scheffé's method for multiple comparisons adjustment in linear models, also use *F*-tests.

In Simple Linear regression model, we are using F test for testing the overall significance of the model. The F-test of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables. The overall F-test compares the model that you specify to the model with no independent variables. This type of model is also known as an intercept- only model.

The F-test for testing the overall significance of the model is build on the following two hypotheses:

- The null hypothesis states that the model with no independent variables fits the data as well as your model.

- The alternative hypothesis says that your model fits the data better than the intercept-only model.

In statistical output, you can find the overall F-test in the ANOVA table. An example is below.

Table 1.3.1 Analysis of Variance

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	12833.9	4278.0	57.87	0.000
East	1	226.3	226.3	3.06	0.092
South	1	2255.1	2255.1	30.51	0.000
North	1	12330.6	12330.6	166.80	0.000
Error	25	1848.1	73.9		
Total	28	14681.9			

Compare the p-value for the F-test to the pre decided significance level. If the p-value is less than the significance level, the sample data provide sufficient evidence to conclude that our regression model fits the data better than the model with no independent variables.

This finding is good because it means that the independent variables in our model proved to be a better fit in the model.

In general, if none of the independent variables are statistically significant, the overall F-test is also not statistically significant. Occasionally, the tests can produce conflicting results. Such problems may creep because the F- test of overall significance assesses all of the coefficients jointly whereas the t-test for each coefficient examines them individually. For example, the overall F-test can find that the coefficients are significant *jointly* while the t-tests can fail to find significance *individually (when individual βs are insignificant based on 't'.)*.

These conflicting test results can be hard to understand, but think about it this way. The F-test sums the predictive power of all independent variables and determines that it is unlikely that *all* of the coefficients equal zero. However, it's possible that each variable isn't predictive enough on its own to be statistically significant. In other words, our sample provides sufficient evidence to conclude that our model is significant, but not enough to conclude that any individual variable is significant.

1.3.6 Standard Error

The standard error of the estimate in regression measures the average distance of observed data points from the regression line. In other words, it tells us how much observed values differ from the predicted values, on average.

It is a crucial measure for assessing the precision of estimated coefficients: smaller standard errors mean more precise estimates. It is used to compute t-statistics for hypothesis testing regarding the regression coefficients

The standard error measures the sampling variability of an estimator. For an OLS coefficient $\hat{\beta}_1$ it answers: "How much would $\hat{\beta}_1$ move across repeated random samples drawn under the same data-generating process?"

If $\hat{\beta}_1$ is the least square estimate of β1, and se($\hat{\beta}_1$) is its standard error. The value of se($\hat{\beta}_1$) can be calculated as follows:

$$se(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

If $\hat{\beta}_0$ is the least square estimate of β0, and se($\hat{\beta}_0$) is its standard error which is calculated using:

$$se(\hat{\beta}_0) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

Illustrative Example

Suppose we regress consumption on income using 500 households.

Simple output table will be as follows:

Table 1.3.2. Output Table

Variable	Coefficient	Std. Error	t-stat	p-vale
Intercept	4.12	1.03	4.01	0.0001
Income	0.78	0.05	15.6	0.0000

Interpretation is that each extra income unit raises consumption by 0.78 units.

Summarised Overview

Ordinary Least Squares (OLS) provides a practical bridge from theory to evidence by selecting, from observed data, the straight line that best summarises the average relationship between a dependent variable and one or more regressors. In a simple linear setting, the sample regression function approximates the population regression function by choosing coefficient estimates that minimise the sum of squared residuals, the vertical gaps between observed and fitted values. Squaring residuals penalises larger deviations more heavily and yields unique estimates for the intercept and slope that make the fitted line pass through the sample means, with residuals that sum to zero and are uncorrelated with both fitted values and the regressor.

Minimising the residual sum of squares proceeds by the differentiation and minimisation principles. Taking partial derivatives of the sum of squared residuals with respect to the coefficients, setting them to zero, and verifying second-order conditions produces the normal equations. Solving these gives the OLS estimators, which are functions only of

observable sample quantities and thus straightforward to compute. OLS and maximum likelihood coincide for the classical normal-error model, but OLS is preferred in many applications for its simplicity and robustness of interpretation.

The usefulness of OLS relies on the assumptions of the Classical Linear Regression Model. The regression must be linear in parameters; the error has zero mean, constant variance, and no serial correlation; there is no perfect multicollinearity among regressors; regressors are non-stochastic (or at least exogenous) with zero covariance with the disturbance; the sample size exceeds the number of parameters; there is genuine variation in the regressors; and the model is correctly specified. Normality of errors is not required for unbiasedness, but it enables exact small-sample inference via t and F distributions. When homoscedasticity and no autocorrelation both hold, the disturbances are independent and identically distributed.

Under these assumptions, the Gauss–Markov theorem shows that OLS is BLUE: among all linear unbiased estimators, the OLS estimators have the smallest variance. Beyond finite-sample properties, OLS is asymptotically unbiased and consistent; as the sample grows, sampling variability shrinks and the estimates concentrate around the true parameters. These properties explain the centrality of OLS in empirical economics, provided that diagnostics confirm the assumptions reasonably hold.

Goodness of fit is summarised by the coefficient of determination, r^2, the proportion of variation in the dependent variable accounted for by the regressors. Algebraically, r^2=ESS/TSS=1−RSS/TSS, where the total variation splits into explained and residual components. Values closer to one indicate that the fitted line tracks the data more closely, though a high r^2r^2r^2 is neither necessary nor sufficient for causal interpretation or predictive validity.

Hypothesis testing evaluates whether estimated relationships are statistically different from hypothesised values. The usual t-test assesses individual coefficients using the estimate divided by its standard error, compared to critical values from the t distribution with appropriate degrees of freedom; a non-rejection of a zero slope implies no detectable linear relation. The overall F-test examines whether the model with regressors improves fit relative to an intercept-only benchmark; it aggregates joint explanatory power and can be significant even when no single coefficient is individually significant. Standard errors quantify sampling uncertainty around estimates and are essential for constructing confidence intervals and test statistics; smaller standard errors imply more precise estimates.

Ordinary Least Squares, then, turns theoretical statements into empirically testable and interpretable results through a transparent optimisation criterion, clear assumptions, and well-developed inferential tools. When assumptions are approximately satisfied—or when appropriate remedies are used for violations—OLS offers estimates that are intuitive, computationally convenient, and statistically reliable for analysis, policy evaluation, and forecasting.

Assignments

1. Explain why OLS minimises the sum of squared residuals and how this relates to maximum-likelihood under normal errors.
2. State and justify each assumption: linearity in parameters, random sampling, full rank X, zero-mean error, homoskedasticity, no autocorrelation, and exogeneity.
3. Prove that, under the classical assumptions, OLS is the Best Linear Unbiased Estimator (BLUE).
4. Compute and interpret R^2 and adjusted R^2 for both cross-section and time-series contexts.
5. Critically assess the limitations of R^2 as a measure of model quality (over-fitting, nonlinearity, out-of-sample performance).
6. Formulate null and alternative hypotheses on single coefficients, linear combinations, and overall model significance.
7. Explain the role of estimated error variance in constructing standard errors.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*): Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

BLOCK 2

Violation of the CLRM Assumptions

UNIT 1

Heteroscedasticity

Learning Outcomes

After completing this unit, the learner will be able to:

- to identify and characterise heteroscedasticity
- evaluate consequences for OLS
- implement appropriate remedies, thereby safeguarding inference validity in applied econometric research

Background

When learning about heteroscedasticity learner need to know what "should" happen under OLS assumptions to understand why non-constant error variance is a problem. They need to grasp how violations affect not the OLS coefficients themselves (which remain unbiased) but the efficiency of estimates and the reliability of standard errors, confidence intervals, and significance tests. Having practiced with model assumptions and diagnostics, learners will better understand the motivation for specific tests and remedies (such as robust standard errors or transforming variables)

Keywords

Homoscedasticity, Park Test, The Breusch - Pegan Test, White test, Spearman's Rank Correlation Test, Glejser Test, GoldFeld - Quandt Test

Discussion

2.1.1 Heteroscedasticity

The classical linear regression model advocates that the disturbances u_i appearing in the population regression function are homoscedastic, which implies they all have the same variance. If the data points are unequally scattered it represents heteroscedasticity. In this lesson we examine to what extent this assumption is valid and find out what happens if this assumption is violated. We seek answers to the following questions:

- What is the nature of heteroscedasticity?
- What are its consequences?
- How can we detect it?
- What are the remedial measures?

2.1.1.1 Nature of Heteroscedasticity

Where the conditional variance of the Y population varies with X, this situation is known appropriately as heteroscedasticity or unequal spread or variance. That is,

$$E(u_i^2) = \sigma_i^2$$

We can illustrate the problem of Heteroscedasticity as in figure given below.

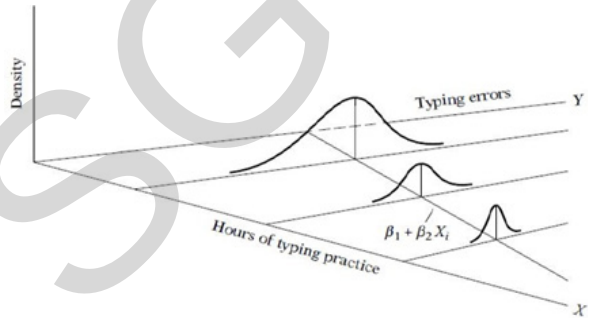


Fig. 2.1.1. Heteroscedasticity

2.1.1.2 Reasons for Heteroscedasticity

Various reasons for the occurrence of Heteroscedasticity are;

1. If the datasets have a large range between the biggest and the smallest observed values, i.e when there are outliers in the dataset.
2. If observations are mixed with different measures of scale.
3. As data collecting techniques increases σ_i^2 is likely to decrease.
4. It can also arise as a result of the presence of collinearity.

5. Skewness in the distribution of one or more regressors included in the model.
6. Incorrect data transformation.
7. Incorrect functional form.
8. Incorrect model specification.

2.1.1.3 Consequences of Heteroscedasticity

In the presence of ***Heteroscedasticity***, we can estimate our regression model and find out the parameters of the model as;

$E(u_i^2)=\sigma_i^2$

$\therefore \quad Y_i = \beta_1 + \beta_2 X_i + u_i$

By using the SRF,

$Yi = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$

Applying the usual formula, the OLS estimator $\hat{\beta}_2$ is $\hat{\beta}_2$ =

$$\frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

$$= \frac{n\Sigma x_i y_i}{n\Sigma X_i} - \frac{\Sigma x_i y_i}{(\Sigma X_i)^2}$$

$$\therefore \quad Var\hat{\beta}_2 = \frac{\sigma_i^2}{\sum x_i^2}$$

Under CLRM, these OLS estimators are BLUE. But with Heteroscedasticity, the consequences are;

1. OLS estimators are still linear
2. OLS estimators are still unbiased
3. OLS estimators no longer have minimum variance. That is, they are no longer efficient. In simple terms, OLS estimators are no longer BLUE in small as well as in large samples.
4. The usual formula to estimate variances of OLS estimators are generally biased. The usual formula is,

$$\therefore \quad Var\hat{\beta}_2 = \frac{\sigma_i^2}{\sum x_i^2}$$

Because of heteroscedasticity we cannot use this formula.

A positive bias occurs if OLS overestimates the true variance of estimator and the negative bias occurs if OLS underestimate the true variance of estimator.

5. The bias arises from the fact that the conventional estimator of true σ^2 is no longer an unbiased estimator of σi^2

6. As a result, the usual confidence intervals and hypothesis tests based on t and F distributions are unreliable. The tests of hypothesis (like t-test, F-test) are no longer valid due to the inconsistency in the co-variance matrix of the estimated regression coefficients. If conventional testing procedures are employed there is a possibility of drawing wrong conclusions

Thus, in the presence of Heteroscedasticity OLS estimators are no longer BLUE. So we have to rely on other methods like Generalized Least Square (GLS) for estimation. Similarly, ordinary testing of hypothesis is not reliable due to the possibility of drawing wrong conclusions. Therefore, it is essential to detect and solve the problem of Heteroscedasticity before estimation.

2.1.1.4 Tests of Heteroscedasticity or Detection of Heteroscedasticity

Tests for identifying hetroscedasticity does not involve any hard and fast rules. We have only certain rules of thump.

Informal Methods

1. **Nature of Problem**: - Nature of the problem under consideration suggests whether Heteroscedasticity is likely to be encountered. Based on the past studies, one can analyse the nature of heteroscedasticity in socio-economic surveys. It is assumed that in similar surveys one can expect unequal variances among the disturbances. Hence, in cross-sectional data involving heterogeneous units, heteroscedasticity may be present.

2. **Graphical Method**: - If there is no empirical information about the nature of Heteroscedasticity, in practice one can do the regression analysis on the assumption that there is no Heteroscedasticity. Then do a post-mortem examination of the residual squared ûi2 to see if they exhibit any systematic pattern. Although ûi2 are not the same thing as ui2, they can be used as proxies especially if the sample size is sufficiently large.

An examination of the $\hat{u}_i^2$ may reveal the following patterns.

Here we are plotting $\hat{u}_i^2$ against the estimated Y values, $\hat{Y}_i$. Then we are finding out whether the $\hat{Y}_i$ is systematically related to $\hat{u}_i^2$. If they show some patterns, it means that there is heteroscedasticity.

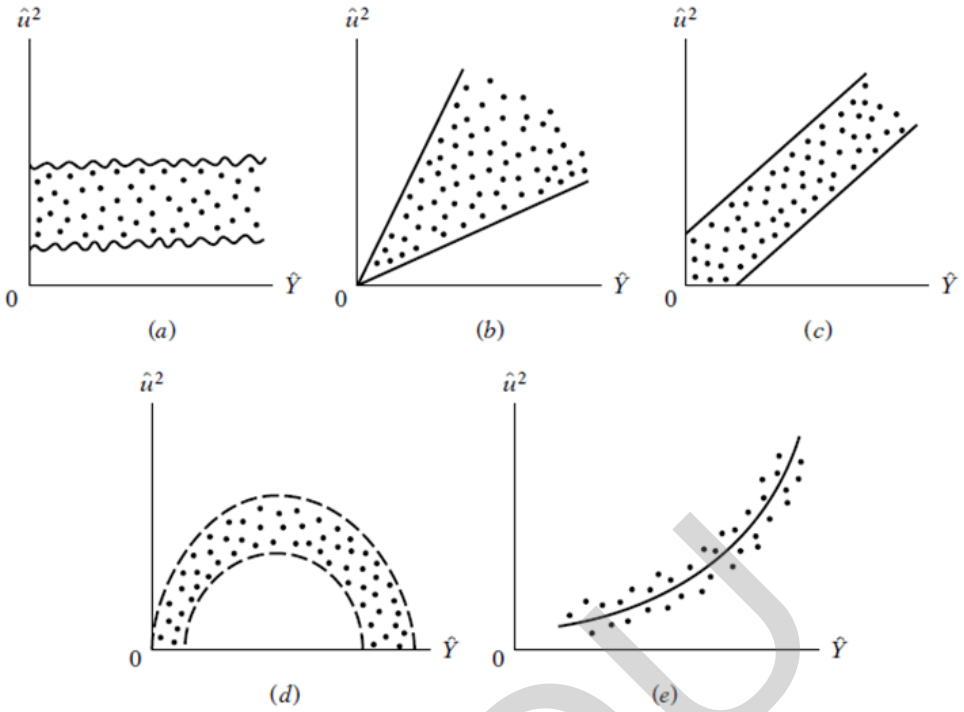


Fig. 2.1.2 Detection of Heteroscedasticity

In figure 'a', we see that there is no systematic pattern between the two variables, suggesting no heteroscedasticity is present in data. But from figures 'b' to 'e' they show some patterns and therefore there is heteroscedasticity in these data.

Formal Methods

1. **Park Test :** - Park formalized the graphical method by suggesting that σi^2 is some function of the explanatory variable Xi.

Suggested functions are

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi}$$
$$or$$
$$\ln\sigma_i^2 = \ln\sigma^2 + \beta \ln X_i + V_i$$

Since σi^2 is generally not known, Park suggested using $\hat{u}_i$, as a proxy and running following regression.

$$\ln\hat{u}_i^2 = \ln\sigma^2 + \beta \ln X_i + V_i$$
$$= \alpha + \beta \ln X_i + V_i$$

If β turn out to be statistically significant, it would suggest that Heteroscedasticity is present in the data.

Park test is a two stage procedure.

a. We run the OLS regression disregarding the heteroscedasticity question.

b. Run the regression

2. The Breusch - Pegan Test: It tests whether the variance of the errors from regression is dependent on the values of the independent variables. In that case, heteroskedasticity is present.

3. White test: White test establishes whether the variance of the errors in a regression model is constant. To test for constant variance one adopts an auxiliary regression analysis. This regresses the squared residuals from the original regression model into a set of regressors that contain the original regressors along with their squares and cross-products.

4. Spearman's Rank Correlation Test:- We know that the Spearman's rank Correlation Coefficient is,

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

Where, di = difference in the rank n= no. of individual

Assume, $Y_i = \beta_0+\beta_1X_i+u_i$

Then the rank correlation coefficient can be used to detect heteroscedasticity as follows:

Step 1 :- fit the regression line to the data on Y and X and obtain the residuals $\hat{u}_i$

Step 2 :- taking their absolute value, $/\hat{u}_i/$, rank both $/\hat{u}_i/$ and X_i or $\hat{Y}_i$ according to an ascending or descending order and compute Spearman's rank correlation coefficient.

Step 3 :- Assuming that the population rank correlation coefficient Ps = 0, and n>8, the significance of the sample rank correlation r_s can be tested by the 't' test as follows,

$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$ with n-2 degrees of freedom.

If the computed 't' value > critical 't' value, we do not reject the hypothesis of heteroscedasticity. Otherwise, if the computed 't' value < critical 't' value, we reject the hypothesis of heteroscedasticity assumption.

If the regression model involves more than one X variable, r_s can be computed between /ûi/ and each X variables separately and can be tested for the statistical significance using 't' test

5. Glejser Test: - Glejser test is similar to Park test in its spirit. After obtaining the residuals $\hat{u}_i$ from the OLS regression, Glejser suggests regressing the absolute values of $\hat{u}_i$ on the X variable that is thought to be closely related with σi^2. Glejser suggested the following functional form for this.

$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + V_i$$

In empirical and practical situations, one can use Glejser approach.

6. GoldFeld - Quandt Test: - One of the popular methods, in which of one assumes that the Heteroscedastic variance σi^2 is positively related to one of the explanatory valuables in the regression model.

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Suppose σi^2 is positively related to Xi as;

$$\sigma i^2 = \sigma i \mathrm{X} i^2$$

Where σ^2 is a constant. This equation gave us the idea that, σi^2 is proportional to the square of the X variable. That is, σi^2 would be larger if X variable become larger. Therefore, heteroscedasticity is more likely to be present in the model.

2.1.1.5 Remedial Measures for Heteroscedasticity

- *When σi^2 is known:* ***The Method of Weighted Least Squares***

This is the generalization of ordinary least square and linear regression in which the errors co-variance matrix is allowed to be different from an identity matrix.

As we have seen, if σi^2 is known, the most straight forward method of correcting heteroscedasticity is by means of weighted least squares. The estimators thus obtained are BLUE. To fit this idea, consider a two variable regression model,

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad \text{.....(1)}$$

Assume that, the true error variance σi^2 is known. That is, the error variance for each observation is known. Now transformation of the model can be done by deflating or dividing both sides of the regression model by the known σ_i

Now let, $v_i = {}^{ui}/_{\sigma_i}$ where, v_i= the transformed error term. If v_i is homoscedastic, then the transformed regression does not suffer from the problem of heteroscedasticity. Thus, it can be estimated using the usual OLS method. Assuming all other assumptions of the CLRM are fulfilled, OLS estimators of the parameters in the equation will be BLUE and we can then proceed to statistical inference in the usual manner.

- *When σi^2 is not known:*

If true σi^2 are known, we can use the WLS method to obtain BLUE estimators. Since the true σi^2 are rarely known. Therefore, if we want to use the method of WLS, we will have to adopt to some adhoc assumption about σi^2 and transform the original regression model so that the transformed model satisfies the homoscedasticity assumption.

- Re-specification of the model

Instead of speculating σi^2, a re-specification of the model choosing a different functional form can reduce heteroscedasticity. For example, instead of running linear regression, if we estimate the model in the log form, it often reduces heteroscedasticity.

Summarised Overview

Heteroscedasticity arises when the variance of the error term in a regression model is not constant across observations, violating one of the Classical Linear Regression Model (CLRM) assumptions. Instead of errors being evenly dispersed (homoscedastic), heteroscedastic data show unequal spread, often illustrated in scatterplots where residuals widen or narrow systematically with fitted values.

The causes include the presence of outliers, mixing units with different scales, collinearity, skewed regressors, poor data transformations, incorrect functional forms, and model misspecification. Its consequences are serious: although OLS estimators remain linear and unbiased, they lose efficiency and are no longer the Best Linear Unbiased Estimators (BLUE). Variance estimates of coefficients become biased, rendering standard errors, confidence intervals, and hypothesis tests unreliable.

Detection methods are informal and formal. Informal approaches include examining the nature of the data (especially cross-sections) and using scatterplots of squared residuals against predicted values. Formal statistical tests include the Park Test, Breusch–Pagan Test, White Test, Spearman's Rank Correlation Test, Glejser Test, and the Goldfeld–Quandt Test. These methods assess whether residual variances systematically depend on explanatory variables.

Remedies include applying Weighted Least Squares (WLS) if the true error variances are known, or adopting feasible WLS when variances are unknown but approximated. Another approach is re-specifying the regression model with alternative functional forms (e.g., log transformations), which often reduce heteroscedasticity. Detecting and correcting heteroscedasticity is essential, since ignoring it undermines statistical inference and may lead to invalid conclusions.

Assignments

1. Distinguish unconditional versus conditional heteroskedasticity and recognize typical empirical contexts.
2. Demonstrate algebraically that heteroskedasticity leaves OLS coefficients unbiased but inflates/deflates their variances.
3. Explain the tests for detecting heteroscedasticity
4. Suggest remedial measures for heteroscedasticity in order to ensure inference validity

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 2

Multicollinearity

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the nature of multicollinearity
- explain the consequences of multicollinearity
- understand the presence of multicollinearity
- implement appropriate remedies for eliminating multicollinearity

Background

The meaning of regression coefficients and how to interpret them in the presence of multiple predictors is essential for learners before understanding the nature of multicollinearity. The importance of variable independence for regression that predictors should ideally supply unique, non-overlapping information about the dependent variable should also be identifiable for the learner. The idea that some problems (like lack of independence among predictors) do not affect the unbiasedness of OLS coefficients, but can affect their precision. How unreliable estimates (due to issues like high correlation among predictors) can lead to misleading or insignificant inferences even if the overall fit of the model seems good should be looked into.

Interpretation of regression output recognizing differences between the statistical significance of individual predictors versus the overall regression model should be possible. Experience in interpreting changes in coefficients when variables are added or dropped from the model is also essential.

Keywords

Multicollinearity, Classical Linear Regression Model, Standard Errors, BLUE, Confidence Intervals, Variables

Discussion

2.2.1 Multicollinearity

Another important assumption of the Classical Linear Regression Model (CLRM) is that there is no Multicollinearity among the regressors included in the multiple regression models. In practice, one rarely encounters perfect multicollinearity but cases of near or very high Multicollinearity can be found, where explanatory variables are linearly correlated in many instances.

The term multicollinearity was coined in 1934 by Ragnar Frisch in his book 'Confluence Analysis'. Because of strong interrelationships among the explanatory variables, it becomes difficult to find out how much each of these will influence the dependent variable. Usually, economic variables are related in several ways and because of inter-relationship among the explanatory variables, often the statistical results gained from them are found to be ambiguous, a multicollinearity problem is said to exist. Under this section, we are explaining the nature, reasons, consequences, detection measures and ways to solve the problem of multicollinearity.

2.2.2 Nature of Multicollinearity

Strictly speaking, Multicollinearity refers to the existence of more than one exact linear relationships and collinearity refers to existence of a single linear relationship. But this distinction is rarely maintained in practice and multicollinearity refers to both cases. It means the existence of a perfect or exact linear relationship among some or all explanatory variables of a regression model. For the k variable regression involving explanatory variables X_1 X_2……. X_k (where X_1 = 1 for all observations to allow for the intercept term), an exact linear relationship is said to exist if the following conditions are satisfied.

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k = 0 \quad .....(1)$$

Where $\lambda_1 + \lambda_2 + ... + \lambda_k$ are constants such that not all of them are zero simultaneously.

But the chances of obtaining a sample of values where the regressors are related in this fashion are rare in practice. Today however the term multicollinearity is used in a broader senseto include the case of perfect multicollinearity (as equation 1) as well as the case where the X variables are inter-correlated but not perfectly so, as follows.

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k + V_i = 0 \quad .....(2)$$

Where, vi=stochastic error term

To understand the difference between perfect and less than perfect multicollinearity in our example assumes that, $\lambda_2 \neq 0$

Then equation (1) can be written as,

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - ..... \frac{\lambda_k}{\lambda_2} X_{k1i} \quad .....(3)$$

The equation (3) shows that X_2 is exactly linearly related to other variables. In this situation the coefficient of correlation between X_2 and the linear combination on the right side of the equation (3) is found to be Unity.

But in the case of less than perfect multicollinearity by assuming $\lambda_2 \neq 0$, equation (2) can also be written as,

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - ..... \frac{\lambda_k}{\lambda_2} X_{k1i} - \frac{1}{\lambda_2} V_i \text{ .....(4)}$$

Equation (4) shows that X2 is not an exact linear combination of other Xs because it is also determined by the stochastic error term vi.

If multicollinearity is perfect, regression coefficients of the X variables are indeterminate and their standard errors are infinite. If multicollinearity is less than perfect, regression coefficients although determinate, but have large standard errors (in relation to the coefficients themselves) which means the coefficients cannot be estimated with much precision or accuracy.

2.2.3 Consequences of Multicollinearity

It can be shown that even if the multicollinearity is very high the OLS estimators are still retain the property of BLUE.

Theoretical Consequences

1. It is true that even in the case of high multicollinearity, the OLS estimators are unbiased, but unbiasedness is a multi -sample or repeated sampling phenomenon. But this says nothing about the properties of estimators in any given sample
2. It is true that collinearity does not destroy the property of minimum variance in the class of all linear unbiased estimators. The OLS estimators have minimum variance that is their efficient. But it does not mean that the variance of OLS estimator will necessarily be small
3. Multicollinearity is essentially a sample phenomenon in the sense that even if the X variables are not linearly related in the population they may be so related in the particular sample.

For these reasons, the fact that the OLS estimators are BLUE despite multicollinearity is of little consolation in practice.

Practical Consequences

1. Although BLUE, the OLS estimators have large variances and co-variances making the precision difficult.
2. Because of this, the confidence intervals tend to be much wider leading to the acceptance of the zero null hypothesis more rapidly.
3. Because of this, the 't' ratio of one or more coefficients tends to be statically insignificant in the case of high collinearity, the estimated standard error increase dramatically by making

t values smaller. Therefore, in such cases one increasingly accepts the null hypothesis

4. Although t ratios of one or more coefficients is insignificant statistically, the R^2 (the overall measure of goodness of fit) can be very high. That is, on the basis of 't' test one or more of the partial slope coefficients are statistically insignificant and we accept H0: $\beta 2 = \beta 3 = \beta k = 0$

   But R^2 is so high, say 0.9, on the basis of F test one can reject H0. That it is one of the signals of multicollinearity - insignificant 't' values, but a high overall R^2 and a significant F value.

5. The OLS estimators and their standard errors can be sensitive to small changes in the data.

2.2.4 Detection of Multicollinearity

Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity but between its various degrees. Since multicollinearity refers to the degree of relationship between explanatory variables that are assumed to be non-stochastic, it is a feature of the sample and not of the population. Since multicollinearity is a sample phenomenon, we do not have one unique method of detecting it for measuring its strength. But we have some rules of thumb which all the same. Some of them are;

1. **High R2 But Few Significant T-Ratios** : This is the Classic symptom of multicollinearity. If R^2 is high ($R^2 > 0.8$), the F test in most cases will reject H0 (H0: β's are zero) but the individual t ratios are insignificant and thus accept H0. Although this diagnostic is sensible, its disadvantage is that it is too strong in the sense that multicollinearity is considered as the harmful only when all of the influence of the explanatory variables on Y cannot be separated.

2. **High Pair Correlations Among Regressors** : Another rule of thumb suggested is that if the pair wise or zero order correlation coefficient between two regressors is high (>0.8) then multicollinearity is a serious problem. It is also clear that high zero-order correlations are a sufficient but not a necessary condition for the existence of multicollinearity because it can exist even the zero-order correlation or simple correlations are comparatively low.

   But in models involving more than two explanatory variables the simple or zero order correlation will not provide an unfailing guide in the presence of multicollinearity. In fact, if there are only two explanatory variables, the zero-order correlation will suffice.

3. **Examination of Partial Correlations :** When there are more than two explanatory variables in the model, we often relied on the partial correlation for detecting multicollinearity. Thus in the regression analysis of Y on X2, X3 and X4, a finding that $R^2_{1.234}$ is very high but $r^2_{12.34}$, r^2 and $r^2_{14.23}$ are comparatively low. It may suggest that the variables X2, X3 and X4 are highly inter-correlated and that at least one of these variables is excessively related with another. But there is no guarantee that the partial correlations will provide an efficient guide to multicollinearity, for it may happen that both R^2 and all the partial correlations are sufficiently high. That is, a given partial correlation may be

compatible with different multicollinearity patterns.

4. **Auxiliary Regressions :** Since multicollinearity arise because one or more of the regressors are in exact or approximately linear combinations with other regressors. One way of finding out which X variable is related to other X variables is to regress each Xi on the remaining X variables and compute the corresponding R^2 (R^2i). Each one of these regressions is called and auxiliary regression. Then the relationship established between F and R^2 in the variable as;

$$R_i^2 = \frac{\left(R_{x1,x2,\ldots xk}\right)^2 /(k-2)}{1-\left(R_{x1,x2,\ldots xk}\right)^2 /(n-k+1)}$$

Follows the F distribution with (k-2) and (n-k+1) degrees of freedom.

In this equation, n = sample size

k = number of explanatory variables

$R^2_{x1, x2...xk}$ = the coefficient of determination in the regression of variable Xi on the remaining X variables.

If computed F value is greater than critical Fi at the chosen level of significance it means that the particular Xi is collinear with other Xs and that variable is dropped from the model

If the computed F is less than the critical Fi we say that it is not collinear with other Xs and retain that variable in the model.

But if there are several complex linear relations this curve fitting exercise may not prove to be of much value as it will be difficult to identify the separate interrelationships

Therefore, one may adopt Klien's rule of thumb instead of testing all auxiliary R^2 values. It is suggested that multicollinearity may be a troublesome problem only if the R^2 obtained from an auxiliary regression is greater than the overall R^2 the one that obtained from the regression of Y on Xs.

5. **Eigen Values and Condition Index**

Eigen values and Condition Index are widely used to detect multicollinearity. From the Eigen values we can derive condition number k as,

k= Maximum Eigen Value/ Minimum Eigen Value.

The Conditional Index (CI) is,

$$CI = \sqrt{K} = \frac{\sqrt{Maximum\ \ Eigen\ Value}}{Mininim\ Eigen\ Value}$$

The rule of thumb for using k and CI for detecting multicollinearity is that,

If k is between 100 and 1000 there is moderate to strong multicollinearity and if k is greater than 1000 there is severe multicollinearity.

Alternatively, if Condition Index is in between 10 and 30 if there is moderate to strong multicollinearity and if Conditional Index is greater than 30 there is severe multicollinearity.

6. **Tolerance and Variance Inflation Factor (VIF):** For k variable regression model (Y, intercept and k–1 regressors) the variance of a partial regression coefficient is,

$$Var(\beta)_j = \frac{(\sigma_2)}{\Sigma X\, j^2} \frac{(1)}{1 - R\, j^2} = \frac{(\sigma^2)}{\Sigma X\, j^2} VIF_j$$

Variance Inflation Factor (VIF) means the speed with which the variance and co-variance increase and it can be expressed as,

$$VIF = \frac{(1)}{1 - r_{23}{}^2} (in \ \ three\ variable\ case)$$

Bj = partial regression coefficient of the regressor Xj

$Rj^2 = R^2$ in the auxiliary regression of the Xj on the remaining (k-2) regressors.

R2 increases towards unity as the collinearity of Xj with other regresses increases, the VIF also increases and in the limit it can be infinite.

Therefore, VIF can be used as an indicator of multicollinearity. The larger the value of VIF the more troublesome or collinear is the variables Xj and vice versa. As a rule of thumb, if the VIF > 10 of a variable that variable is set to be highly collinear

Tolerance can also be used to detect multicollinearity. It is defined as,

$TOLj = (1-Rj^2)$

= 1- VIFj

TOLj = 1, If Xj is not correlated with other regressors

TOLj =0 If Xj is perfectly related to other regressors.

2.2.5 Remedial Measures

Elimination of the effect of multicollinearity is not an easy task. There is no sure-fire remedy, but there are only a few rules of thumb because it is a sample phenomenon. Besides despite near collinearity, OLS estimators still retain their BLUE property. The following are the solutions for the incidence of multicollinearity.

1. **A-Priory Information:** It is possible that we can have some knowledge of the values of one or more parameters from previous empirical work. This knowledge can be profitably utilised in the current sample to reduce multicollinearity.

2. **Combining Cross Sectional and Time Series Data:** Another technique to reduce the effect of multicollinearity is to combine cross-sectional and time series data, that is, pooling the data.

3. **Dropping A Variable(S) And Specification Bias:** One simplest method when faced with severe multicollinearity is that to drop one of the collinear variables. Then the model becomes highly significant. But dropping a variable from the model to alleviate the problem of multicollinearity may lead to the specification bias. Hence the remedy may be worse than the disease in some situations because whereas multicollinearity may prevent precise estimation of the parameters of the model, omitting a variable may seriously mislead us as to the true values of the parameters. The OLS estimators are BLUE despite near linearity.

4. **Transformation of the Variables:** In Economics, we have time series data and we know that one reason for high multicollinearity between economic variables is that over time these variables tend to move in the same direction. Therefore, the transformation of the model can minimise if not solve the problem of collinearity. Commonly used transformation technique is first difference form.

5. **Additional or New Data:** Multicollinearity is a sample feature not a population problem it is possible that another sample involving the same variables collinearity may not be so serious. Sometimes simply increasing the sample size may reduce the collinearity problem. The larger the data set, the more the variations in the series that can be captured.

6. **Re-Thinking the Model:** Sometimes a model chosen for empirical analysis is not carefully thought out. Sometime some important variables may be omitted or may be the functional form of the model is incorrectly chosen. However, a proper specification of the model may reduce the problem of multicollinearity.

7. **Other Remedies**

Other remedies for multicollinearity are;

- Factor analysis,
- Principal component analysis, and
- Ridge regression etc.

Summarised Overview

Multicollinearity occurs when explanatory variables in a multiple regression model are highly or perfectly correlated, violating an assumption of the Classical Linear Regression Model (CLRM). Ragnar Frisch coined the term in 1934 in *Confluence Analysis*. In practice, perfect multicollinearity is rare, but high or near-perfect multicollinearity is common, especially in economics, where variables are naturally interrelated. This interdependence makes it difficult to isolate the effect of individual variables on the dependent variable, leading to ambiguous or unreliable statistical results.

The nature of multicollinearity distinguishes between perfect and less-than-perfect cases. In perfect multicollinearity, one explanatory variable is an exact linear combination of others, making coefficient estimation impossible as standard errors become infinite. In less-than-perfect multicollinearity, coefficients can still be estimated, but their standard errors are excessively large, reducing precision and reliability.

Although OLS estimators remain BLUE even with multicollinearity, the consequences are severe in practice. Theoretical implications include unbiased but imprecise estimates, while practical issues involve inflated variances, wide confidence intervals, statistically insignificant t-tests, but simultaneously high R^2 values and significant F-tests. This paradox—low significance of individual coefficients but high overall fit—is a key signal of multicollinearity. Moreover, results become highly sensitive to small data changes, further undermining credibility.

Detection methods include several diagnostic rules of thumb. Common indicators are high R^2 with insignificant t-ratios, pairwise correlations above 0.8, and examination of partial correlations. More formal approaches involve auxiliary regressions and Klein's rule, Eigenvalues and Condition Index, and tolerance/Variance Inflation Factor (VIF), with VIF values above 10 indicating serious multicollinearity. These tests assess the degree rather than the presence of the problem, as multicollinearity is a sample phenomenon.

Remedial measures are varied but imperfect. Options include incorporating prior information from earlier studies, pooling cross-sectional and time-series data, or dropping collinear variables (though this risks specification bias). Transforming variables, such as taking first differences in time series, or expanding sample size, can mitigate the issue. More sophisticated approaches include factor analysis, principal component analysis, and ridge regression, which adjust estimation to reduce multicollinearity's adverse effects. Ultimately, careful model specification and thoughtful data handling are essential to minimising the problem.

Assignments

1. Demonstrate algebraically how multicollinearity inflates the variance of OLS estimators.
2. Discuss the detection methods and suggest remedial measures for multicollinearity.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*): Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 3

Autocorrelation

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the existence of autocorrelation
- elimination of autocorrelation for forecasting accuracy and policy simulation

Background

To effectively learn about Autocorrelation—its nature, consequences, detection, and remedial measures, students should first have a solid understanding of several foundational concepts from regression analysis and econometrics. Familiarity with classical OLS assumptions, especially the assumption that error terms (residuals) are independent and identically distributed and knowing what residuals/errors represent and why their properties matter for valid statistical inference. Moreover, awareness that time series observations can be related to preceding or succeeding values—an idea that sets the stage for understanding autocorrelation. Ability to generate and interpret residual plots as a check for patterns, including non-constant variance or serial patterns is also essential.

By building on these foundations, students will be equipped to understand what autocorrelation is, why it matters in time series (and sometimes cross-sectional) analysis, how to test for and diagnose it, and which corrective techniques are available to ensure valid and reliable econometric modeling.

Keywords

Autocorrelation, Cross Section, Time Series, Pooled Data, Spatial Autocorrelation, Positive Autocorrelation, Negative Autocorrelation, Lagged Variables, Data Transformation, OLS Estimators, BLUE, Runs Test, Durbin–Watson Statistic

Discussion

2.3.1 Autocorrelation

There are generally three types of data that are available for empirical analysis:

1. Cross section,
2. Time series, and
3. Combination of cross section and time series, also known as pooled data.

In developing the classical linear regression model (CLRM) we made several assumptions. However, we noted that not all these assumptions would hold in every type of data. As a matter of fact, we saw in the previous section that the assumption of homoscedasticity, or equal error variance, may not be always tenable in cross-sectional data. In other words, cross-sectional data are often plagued by the problem of heteroscedasticity.

However, in cross-section studies, data are often collected on the basis of a random sample of cross-sectional units, such as households (in a consumption function analysis) or firms (in an investment study analysis) so that there is no prior reason to believe that the error term pertaining to sample is correlated with the error term of another sample. If by chance such a correlation is observed in cross-sectional units, it is called spatial autocorrelation, that is, correlation in space rather than over time.

However, it is important to remember that, in cross-sectional analysis, the ordering of the data must have some logic, or economic interest, to make sense of any determination of whether (spatial) autocorrelation is present or not. The situation, however, is likely to be very different if we are dealing with time series data, for the observations in such data follow a natural ordering over time so that successive observations are likely to exhibit inter-correlations, especially if the time interval between successive observations is short, such as a day, a week, or a month rather than a year. If you observe stock price indexes it is not unusual to find that these indices move up or down for several days in succession. Obviously, in situations like this, the assumption of no auto, or serial, correlation in the error terms that underlies the CLRM will be violated. This situation is termed as the autocorrelation.

Here we are interested to explain,

- The nature of autocorrelation
- The reasons for autocorrelation
- Theoretical and practical consequences of autocorrelation
- The measures to detect the problem of autocorrelation and
- The measures to solve autocorrelation

2.3.2 Nature of Autocorrelation

If there are no correlation between members of series of observation ordered in time (as in time series data) or space (as in cross-sectional data) is known as the assumption of no autocorrelation. That is,

Autocorrelation doesn't exist in the disturbance ui if, E(ui, uj) = 0, if i≠j

Otherwise, if the disturbance terms of a dataset that are ordered in time or space are correlated each other, the situation is generally termed as autocorrelation. That is,

$$E(ui, uj) \neq 0, \text{ if } i \neq j$$

Now let us see some possible patterns of auto and no autocorrelation as figure given below.

On the vertical axis of the figure given below, we take both population disturbances (u) and its sample counterpart (û) and on the horizontal axis time. Then we plot the corresponding points.

From the figure given below, Part (a) to Part (d) errors follow some systematic patterns. Hence, there is autocorrelation. Bute Part (e) reveals no such patterns and hence there is no autocorrelation.

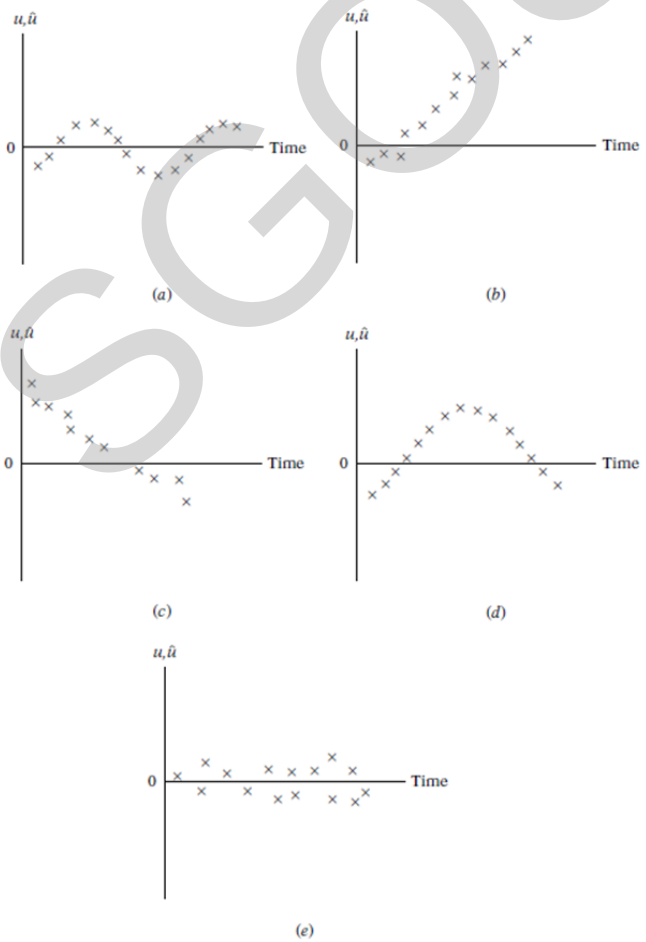


Fig.2.3.1. Patterns of Autocorrelation

Positive and Negative autocorrelation

Autocorrelation can be positive or negative. The value of autocorrelation varies from -1 (for perfectly negative autocorrelation) and 1 (for perfectly positive autocorrelation). The value closer to 0 is referred to as no autocorrelation.

Positive autocorrelation occurs when an error of a given sign between two values of time series lagged by k followed by an error of the same sign. When data exhibiting positive autocorrelation is plotted, the points appear in a smooth snake- like curve, as on the left in figure given below.

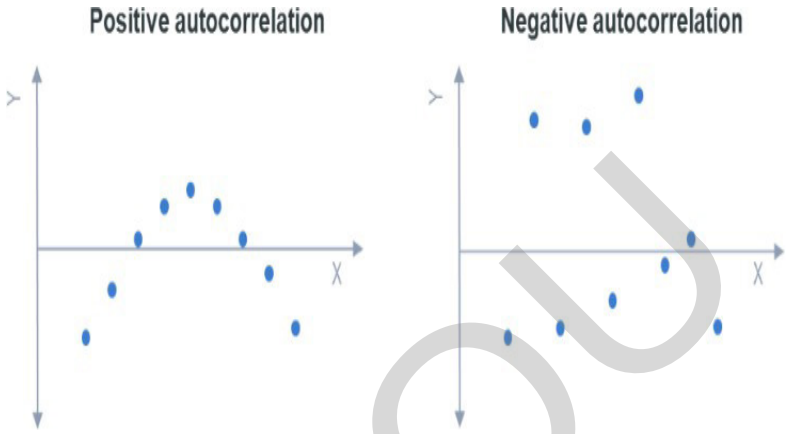


Fig.2.3.2. Correlation

Negative autocorrelation occurs when an error of a given sign between two values of time series lagged by k followed by an error of the different sign. With negative autocorrelation, the points form a zigzag pattern if connected, as shown on the right of figure given above.

2.3.2.1 Reasons for Autocorrelation

The following are the major reasons for autocorrelation.

1. *Inertia:* - Silent feature of most of the time series is inertia or sluggishness. Well known examples in time series are GNI, price Index.
2. *Specification Bias:* Excluded variable case: - Residuals (which are estimate of ui) may suggest that same variable that were originally candidates but were not included in the model for a variety of reasons should be included.

$$Y_i = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_i$$

Yi = Quantity of beef demanded.

X2 = Price of beef

X3 = Consumer income X4 = Price of Pork

t = Time After Regression,

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + V_t$$

3.*Specification Bias:* Incorrect functional form:-

For explaining this, first we are taking case of a marginal cost function,

Marginal Costt = $\beta_1 + \beta_2$ output + $\beta_3 \text{output}_i^2 + u_i$

But instead of this, suppose we get the following model.

$$MC_t = \alpha_1 + \alpha_2 \text{ output}_t + V_i$$

This can be depicted as figure given below.

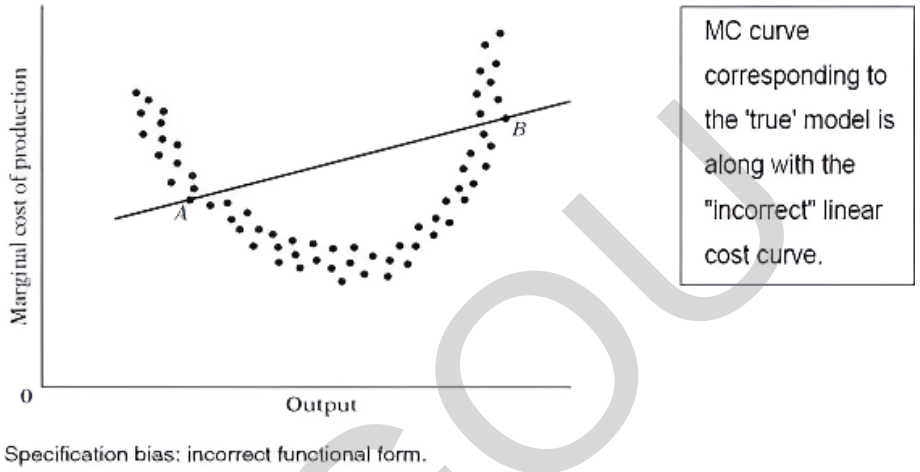


Fig 2.3.3 MC Curve

4.*Cobweb Phenomenon:* - The supply of many agricultural commodities reflects the socalled cobweb Phenomenon.

Where supply reacts to price with a lag of one time period because supply decisions take time implement.

$$Supply_t = \beta_1 + \beta_2 P_{t-1} + u_t$$

1. Lag: - In time series regression model, sometimes the lagged value of the dependent variable also included as one of the explanatory variable. For example,
2. Consumption= $\beta_1 + \beta_2$ *Income* + β_2 $Consumption_{t-1} + u_t$
3. Manipulation of data: - In empirical analysis the raw data are often manipulated.

Data Transformation:- Sometimes data transformation leads to autocorrelation. For example,

$$Y_t = \beta_1 + \beta_2 X_t + u_t \rightarrow I$$

Y = Consumption, X = Income

$Y_{(t-1)} = \beta_1 + \beta_2 X_{(t-1)} + u_{(t-1)}$ → 2 Previous Period

$Y_{(t-1)}$, $X_{(t-1)}$, $u_{(t-1)}$ are lagged values of X_1 Y & U

Sub. (II) from (I) we get

$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t$ → Δ first difference operator

For empirical purpose, $\Delta Y_t = \beta_2 \Delta X_t + V_t \rightarrow V_t = \Delta u_t = (u_t - u_{t-1})$

2.3.3 Consequences

In the presence of autocorrelation, one should not use OLS for estimation, to establish confidence intervals and to test hypothesis. We should use Generalised Least Squares (GLS) method for these purposes. Because in the presence of autocorrelation,

1. The least square estimators are still linear and unbiased.
2. But they are not efficient compared to the procedures that take into account autocorrelation. In short, the usual OLS estimators are not BLUE because they do not possess the property of minimum variance.
3. The estimated variances of OLS estimators are biased. Sometimes, the usual formulas to compute the variances and standard errors of OLS estimators seriously underestimate true variances and standard errors, there by inflating _t' values
4. Therefore, the usual _t' and F tests are not generally reliable.
5. The usual formula to compute the error variance is a biased estimator of true σ^2 .
6. As a consequence, the conventionally computed R^2 may be unreliable measure of true R^2

The conventionally computed variances and SEs of forecast may also be inefficient.

2.3.4 Detection Measures

There are varieties of tests to detect autocorrelation.

1. Graphical Method

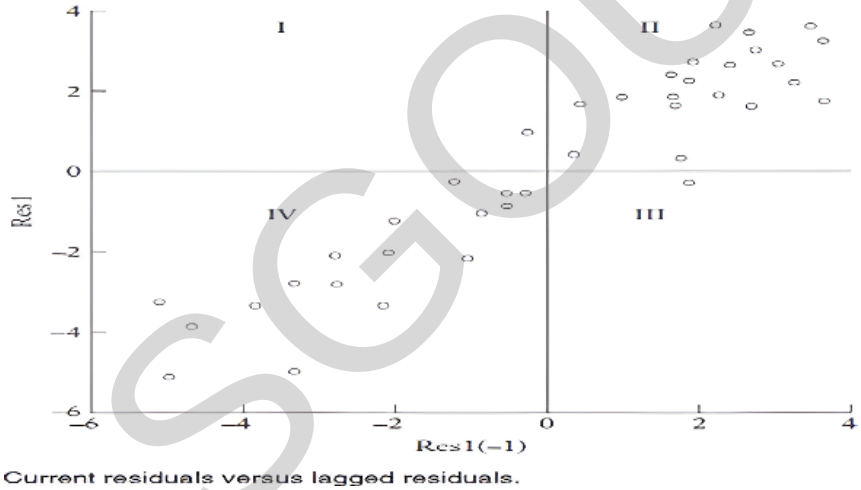
There are various ways of examine the residuals (error) under graphical method.

a) Time Sequence Plot

Residuals and standardized residuals from the wages–productivity regression (

Fig. 2.3.4 Time Sequence Plot

b) Standardized Residual



Current residuals versus lagged residuals.

Fig. 2.3.5. Standardized Residual

Both figures given above clearly shows that the residuals follow some systematic patterns and hence there is autocorrelation.

2. Runs Test

Initially, we have several residuals that are negative, then there is a series of positive residuals, and then there are several residuals that are negative. If these residuals were purely random, could we observe such a pattern? Intuitively, it seems unlikely. This intuition can be checked by the so-called runs test, sometimes also known as the Geary test, a nonparametric test. This is also a crude method.

For the Runs test, let us simply note down the signs of the residuals as+ or -. Suppose we have these signs as;

(---------) (++++++++++++++++++++) (---------)

We now define a run as an uninterrupted sequence of one symbol or attribute, such as + or -. We further define the length of a run as the number of elements in it.

By examining how the runs behave in a strictly random sequence of observations, we can derive a test of randomness of runs. If there are too many runs, it means that the residuals change sign frequently, thus suggesting negative autocorrelation. Similarly, if there are too few runs, it suggests positive autocorrelation.

3. Durbin -Watson 'd' Statistic

It is one of the good methods as the d statistic is based on the estimated residuals, which are computed in regression analysis. It is defined as;

$$d = \frac{\Sigma(\hat{u}_t - \hat{u}_{t-1})^2}{\Sigma(\hat{u}_2)^2}$$

It is simply the ratio of the sum of squared differences in successive residuals to the RSS. It is note that in the 'd' statistic, the number of observations is n -1 because one of the observation is lost in taking successive differences.

A great advantage of this statistic is that it is based on the estimated residuals, which are routinely computed in regression analysis. Because of this advantage it is now a common practice to report the Durbin Watson'd' along with summary statistics such as R^2, adjusted R^2, t ratio etc.

Durbin Watson 'd' statistic is based on some assumptions as

- The regression model includes an intercept term
- The explanatory variables the Xs are not stochastic or fixed in repeated sampling
- The disturbances Ut are generated by the first order autoregressive scheme
- The regression model does not include lagged values of the dependent variable as one of the explanatory variables
- There are no missing observations in the data

He expanded the formula of d statistic as follows;

$$d = \frac{\Sigma\hat{u}_t^2}{\Sigma\hat{u}_t^2} + \frac{\Sigma\hat{u}_{t-1}^2}{\Sigma\hat{u}_t^2} - \frac{2\Sigma\hat{u}_t + \hat{u}_{t-1}^2}{\Sigma\hat{u}_t^2}$$

Since, $\Sigma\hat{u}_t^2$ and $\Sigma\hat{u}_{t-1}^2$ differ in only one observation they are approximately equal. Therefore setting

$\Sigma\hat{u}_t^2 = \Sigma\hat{u}_{t-1}^2$ may be written as

$$d \cong 2\frac{\Sigma\hat{u}_t^2 - \Sigma\hat{u}_{t-1}^2}{\Sigma\hat{u}_t^2}$$

Now let us define the coefficient of autocorrelation, ρ, which can be determined with the help of the sample first-order coefficient of autocorrelation, $\hat{\rho}$

$$\hat{\rho} = \frac{\Sigma\hat{u}_t - \hat{u}_{t-1}}{\Sigma\hat{u}_t}$$

The d statistic become ;

$$d \simeq 2(1-\hat{\rho})$$

Since the value of ρ lies between -1 and + 1 it implies that the value of ‗d' lies between 0 and 4. That is,

d will be $0 \leq d \leq 4$

because $\rho = -1 \leq \rho \leq 1$

→ $d \simeq 2 \rightarrow$ no autocorrelation

→ d ≃ 0 or 4 (closer) there is autocorrelation

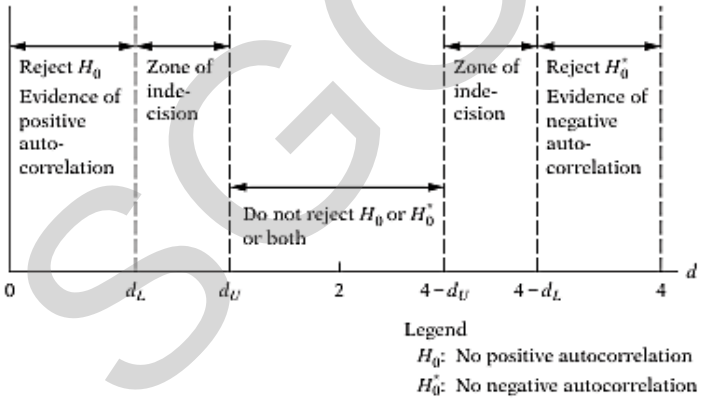


Fig. 2.3.6 Durbin Watson Autocorrelation

2.3.5 Remedial Measures

1. Try to find out if the autocorrelation is pure autocorrelation or not because of the result of the mis- specification of the model.
2. Transformation of original model, so that in the transformed model we do not have the problem of (Pure) autocorrelation.
3. In case of large sample we can use Newey-West method to obtain standard error of OLS estimators that are corrected for auto correlation.
4. In some situation we can continue to use the OLS method.

Summarised Overview

Autocorrelation occurs when error terms in a regression model are correlated across time or space, violating the CLRM assumption of error independence. While cross-sectional data are often affected by heteroscedasticity, time series data are more prone to autocorrelation because observations follow a natural order over time. Positive autocorrelation arises when errors of the same sign persist across periods, producing smooth patterns, while negative autocorrelation alternates error signs, creating zigzag patterns. Spatial autocorrelation can also occur in cross-sectional data when errors across units are interrelated.

The main causes include inertia in time series data (e.g., national income, price indices), specification bias due to omitted variables or incorrect functional forms, lag structures such as the cobweb phenomenon in agriculture, inclusion of lagged dependent variables, manipulation or transformation of data, and delays in adjustment processes.

Although OLS estimators remain linear and unbiased in the presence of autocorrelation, they lose efficiency and are no longer BLUE. Variances and standard errors may be biased, leading to unreliable t and F tests, distorted confidence intervals, and misleading R^2 values. Forecasts also become inefficient.

Detection methods include graphical approaches such as time-sequence plots and residual analysis, the runs test (which checks for randomness in the sequence of residual signs), and the widely used Durbin–Watson statistic. The Durbin–Watson test evaluates first-order autocorrelation, with values close to 2 suggesting no autocorrelation, values approaching 0 indicating strong positive autocorrelation, and values nearing 4 implying strong negative autocorrelation.

Remedies depend on the source of the problem. If autocorrelation is due to model misspecification, re-specifying the functional form or including omitted variables may solve it. Transformations of the regression model, or use of Generalised Least Squares (GLS), are common approaches. For large samples, the Newey–West method provides corrected standard errors. In some cases, OLS may still be usable with caution, but addressing autocorrelation improves reliability of inference and forecasting.

Assignments

1. Diagnose autocorrelation in an empirical macro-time-series model, quantify its severity with appropriate tests, apply a suitable correction, and clearly communicate how the remedy alters coefficient precision, hypothesis tests, and policy conclusions.
2. Explain the concept of autocorrelation in the context of regression analysis. Discuss its nature, distinguishing between spatial autocorrelation in cross-sectional data and serial autocorrelation in time series data.
3. Discuss the major reasons for the occurrence of autocorrelation in regression models.
4. Examine the theoretical and practical consequences of autocorrelation on the Ordinary Least Squares (OLS) method of estimation.
5. Critically evaluate the different methods available to detect autocorrelation in regression analysis.
6. Suggest and explain various remedial measures to address the problem of autocorrelation in regression models.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 4 Model Specification and Errors

Learning Outcomes

After completing this unit, the learner will be able to:

- clearly define a correctly specified econometric model and enumerate the main types of specification errors
- understand the consequences of specifying models incorrectly
- detect potential measurement error and apply an appropriate correction

Background

To successfully study the topics of Model Specification and Errors, their Consequences, Underfitting and Overfitting, and Measurement Errors, students should have the following background knowledge from prior coursework in statistics, econometrics, and basic data analysis:

- The interpretation of regression outputs (e.g., coefficients, R-squared).
- Why OLS assumptions matter and how violations affect coefficient estimation and inference.
- Familiarity with model selection criteria (such as the theoretical justification for including variables).
- Distinguishing between measurement error (inaccurate data) and model error (inadequate model).
- Understanding how errors in the inputs (independent variables) can bias estimates and reduce model reliability.
- Experiences with Underfitting and Overfitting

Keywords

Specification Errors, Under fitting, Over-fitting, Durbin Watson 'd' Statistic. Ramsey's RESET Test

Discussion

2.4.1 Model Specification and Errors

One important assumption of Classical Linear Regression Model is that the regression model is correctly specified or there is no specification bias in the chosen regression model. With this assumption we are estimating the parameters of the chosen regression model and testing hypothesis about them using R^2, 'F', 't', etc. If the tests are satisfactory, the regression model is considered as best fit. If the tests are unsatisfactory, there are some specification errors or bias in the chosen model, such as;

- Whether some important variables are omitted from the model?
- Whether some superfluous (excessive) variables included in the model?
- Is the functional form of the chosen model correct?
- Is the specification of the stochastic error correct?
- Is there more than one specification error?

If these kinds of specification errors are there, the traditional econometric methodology used is Average Economic Regression (AER).

If for example, the bias results from omission of variables, the researcher starts adding new variables to the model and tries to 'build up' the model. This traditional approach to econometric modelling is called the 'bottom-up' approach because we start our model with a given number of regressors and based on diagnostics; go on adding more variables to the model. This approach is also known as Average Economic Regression (AER)

Even though so many criticisms have been raised against the Average Economic Regression, it is still having a place in the standard methodology. Here we are analysing various specification bias and how average economic regression handles the various kinds of specification errors. Before that, we are analysing how average economic regression methodology chooses a regression model first. For this, it uses the following criteria.

Parsimony: - A model can never be a completely accurate description of reality. To describe the reality, one may have to develop such a complex model that will be of little practical use. Some amount (degree) of abstractions for simplification is inevitable in any model building. The principle of parsimony states that, a model be kept as simple as possible. This means that one should introduce in the model a few key variables that capture the essence of the phenomenon under study and retain all minor and random influences to the error term u_i.

Identifiability: - For a given set of data, the identifiability means that the estimated parameters must have unique values. Or what amounts to the same thing, there is only one estimate for a given parameter.

Goodness of Fit: - Since the basic thrust of regression modelling is to explain as much of the variations in the dependent variable as possible by the explanatory variables included in the model. A model is judged good if this explanation, as measured by R^2 is high as possible.

Theoretical Consistency: - A model may not be good, despite a high R^2 if one or more of the estimated coefficients have wrong signs. If for example, in the demand function if one were to obtain a positive sign for the coefficient of the price (positively sloped demand curve) one should look at that result with great suspicion. Therefore, theoretical consistency should be there when framing the models.

Predictive Power: - The only relevant test of the validity of a model is comparison of its prediction with the experience. A high R^2 is used to show the predictive power of the model within the given sample. But we want is its predictive power outside the sample period.

2.4.1.1 Types of Specification Errors

Assume that based on the theory and empirical literature, we accept a good model and let the model is,

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{1i} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where

Y = total cost of production X = output

But suppose that for some reason a researcher decided to use the following model;

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2 + u_{2i} \quad \ldots\ldots\ldots\ldots (2)$$

Since (1) is assumed true, adopting (2) would constitute a specification error, the error consisting in omitting an irrelevant variable X_i^3. Therefore, the error term u_{2i} in (2) is in fact,

$$u_{2i} = u_{1i} + \beta_4 X_i^3 \ldots\ldots\ldots\ldots (3)$$

Now suppose another researcher uses the model,

$$Y_i = \lambda_1 + \lambda_2 X_i + \lambda_3 X_i^2 + \lambda_4 X_5^3 \lambda X^4 + u_{3i} \quad \ldots\ldots\ldots\ldots (4)$$

If (1) is the correct, model (4) also constitutes a specification error, the error here consisting in including an unnecessary or irrelevant variable X_i^4. The new error term is in fact

$$U_{3i} = u_{\$i} - \lambda X^4 \ldots\ldots\ldots\ldots (5)$$

$$= u_{1i} \text{ since, } \lambda_5 X_i^4 = 0.$$

Now assume that the model used is,

$$\ln Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 X_i^2 + \gamma_3 X_i^3 + u_{4i} \quad \text{.....(6)}$$

In relation to the true model (1), the model (6) would also constitute a specification bias, the bias here being the use of the wrong functional form. In (1) Y appears linearly were as in (6) it appears log linearly.

Finally consider another model,

$$Y^*_i = \beta^*_1 + \beta^*_2 X^*_i + \beta^*_3 X^{*2}_i + \beta^*_4 X^{*3}_i + u^*_{1i} \quad \text{.....(7)}$$

Where;

$Y^*_i = Y_i + \epsilon_i$ and

$X^*_i = X_i + w_i$ and

ϵ_i and w_i = errors of measurement .....(7)

states that instead of using the true Yi and Xi we use their

proxies Y*and X* which may contain _errors of measurement

bias‘. To sum up, having once specified a model as the correct model one is likely to commit one or more of these specification errors:

1. Omission of a relevant variable
2. Inclusion of an unnecessary variable
3. Adopting the wrong functional form
4. Errors of measurement

Finally there is one more specification error which is most important it is

- Model misspecification error

This error occurs because we do not know what is the true model in the first place.

2.4.1.2 Consequences of Specification Errors

Whatever be the source of specification error, its consequences are very important. Here we are explaining two kinds of specification errors in the case of three variable regression models and this can be generalized to k-variable case.

- Omitting a relevant variable (under-fitting a model) and
- Inclusion of an irrelevant variable (over-fitting a model)

2.4.1.3 Under fitting a Model (Omitting a relevant variable)

Assume the true model is,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (1)$$

But for some reasons, we fit the following model;

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i \qquad (2)$$

The consequences of omitting X_3 are as follows:

1. If the left out variable X_3 is correlated with the included variable X_2, $r_{23} \neq 0$, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are biased as well as inconsistent. That is, $E(\hat{\alpha}_1) \neq \beta_1$ and $E(\hat{\alpha}_2) \neq \beta_2$ This bias does not disappear even in large samples.
2. Even if X_2 and X_3 are uncorrelated, that is, $r_{23} = 0$, $\hat{\alpha}_1$ is still biased, although $\hat{\alpha}_2$ is unbiased.
3. The var $(\hat{u}_i) = \sigma^2$ is incorrectly estimated
4. Var($\hat{\alpha}_2$) = $\sigma^2 / \sum X^2$ $var(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_i^2}$ is a biased estimator of the variance of $\hat{\beta}_2$
5. The usual confidence interval and hypothesis testing procedures are likely to give misleading conclusions about the statistical significance of the estimated parameters

2.4.1.4 Over-fitting a Model (Inclusion of an irrelevant Variable)

Assume the true model is,

$Y_i = \beta_1 + \beta_2 X_{2i} + u_i$ .....(1)

After committing the specification due to the inclusion of an unnecessary variable the model is,

$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i$ .....(2)

The consequences of this specification error are as follows;

1. The OLS estimators of the parameters of the incorrect model are all unbiased and consistent. That is,

   $E(\hat{\alpha}_1) = \beta_1$, $E(\hat{\alpha}_2) = \beta_2$ and $E(\hat{\alpha}_3) = \beta_3$
2. The error variance σ^2 is correctly estimated
3. The usual confidence interval and hypothesis testing procedures remain valid
4. The estimated α's will be generally inefficient that is their variances will be generally larger than those of the of the true model.

2.4.2 Tests of Specification Errors

Once we identify that there are specification errors, there are remedies for that. Therefore, it is essential to detect whether there is any specification errors in the fitted regression model. The detection measures of specification errors are as follows:

Detecting the Presence of Unnecessary Variables

Suppose we develop a k variable model to explain a phenomenon:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots + \beta_k X_{ki} + u_i \ldots\ldots\ldots\ldots\ldots (1)$$

Suppose we are not sure that X_k is really belongs there, the simple way to find this is to test the significance of the

estimated β_k with the usual‗t' test $t = \hat{\beta}_k / se(\beta_k)$But suppose that we are not sure whether X_3 and X_4 legitimately belong in the model. In this case, we would like to test whether $\beta_3 = \beta_4 = 0$. This can be easily accomplished by the F test. Thus dictating the presence of an irrelevant variable is not a difficult task.

Tests for Omitted Variables and Incorrect Functional Form

To determine whether there is any specification bias due to omitted variables or wrong functional form the commonly used test are;

1. Examination of Residuals

Like autocorrelation and heteroscedasticity, the specification errors due to omission of a relevant variable and wrong functional form can also be detected by examining the residuals. Here also the residuals, if we plot, exhibit distinct patterns

Suppose we have a total cost function:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_{i\,4}^{\,2} + {}_i\beta\, X^3 + u_{1i} \quad \ldots\ldots\ldots (1)$$

Where

Y = total cost of production X = output

But if the researcher fits the model,

$$Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^{\,2} + u_{2i} \quad \ldots\ldots\ldots (2)$$

And another researcher fits the model, $Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \ldots\ldots\ldots (3)$

The (2) and (3) have specification errors. If we plot the residuals we may have as figure 2.4.1.

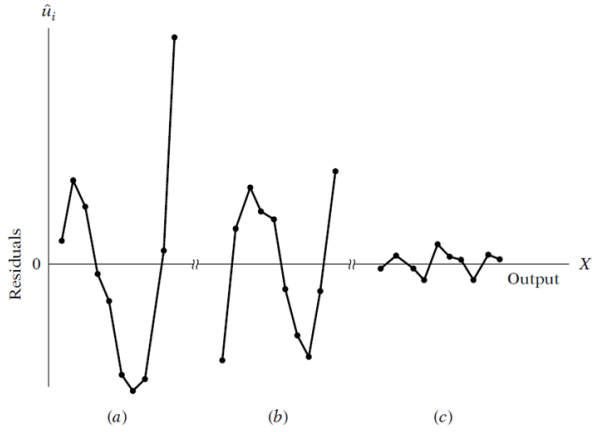


Fig. 2.4.1. Examination of Residuals

In the Figure 2.4.1 as we move from left it to right (a to b to c) the residuals are not only true but also they do not exhibit the pronounced cyclical swings associated with the mis-fitted models. Therefore, if there are specification errors, the residuals will exhibit noticeable patterns.

2. The Durbin Watson 'd' Statistic

The following steps are required for dictating specification errors using Durbin- Watson 'd' statistic.

- From the assumed model, obtain OLS residuals.
- It is assumed that this model is

  $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + u_{1i}$ ..... (1)

  because it excludes at relevant explanatory variable say Z. Therefore, order the obtained residuals from step 1 according to the increasing values of Z.
- Compute the 'd' statistic from the residuals thus ordered by the usual 'd' formula, $d = \frac{\sum(\hat{u}_t - \hat{u}_{t-1})^2}{\sum(\hat{u}_t)^2}$
- Base on Durbin - Watson table, if the estimated 'd' value is significant, and then one can accept the hypothesis of model misspecification. Otherwise reject the hypothesis.

3. Ramsey's RESET Test

Ramsey has proposed a general test of specification error called RESET (Regression Specification Error Test). Let us assume that the SLRM,

$$Y_i = \lambda_1 + \lambda_2 X_i + u_{3i} \ldots\ldots\ldots\ldots\ldots (1)$$

To detect the specification error in the model, the steps involved in RESET are,

- From the chosen model obtain $\hat{Y}_i$ and $\hat{u}_i$.
- Plot the $\hat{u}_i$ in the graph paper to observe whether they exhibit any noticeable pattern
- If the $\hat{u}_i$ are distributed to exhibit some pattern, re-run the regression in introducing $\hat{Y}_i$ in some form as an additional regressor such as $\hat{Y}_i^2$ or $\hat{Y}_i^3$ etc. Thus, we run

  $Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i \ldots\ldots\ldots (2)$
- Obtain R^2 from (1) and (2) as R^2_{old} and R^2_{new}. Then using the F test the statistical significance of increase in R^2 using,

$$F = \frac{\frac{(R^2_{new} - R^2_{old})}{(number\ of\ new\ regressors)}}{\frac{(1 - R^2_{new})}{(n - number\ of\ parameters\ in\ the\ new\ model)}}$$

Summarised Overview

Model specification is a fundamental assumption of the Classical Linear Regression Model (CLRM). A correctly specified regression model should include all relevant variables, exclude irrelevant ones, adopt the appropriate functional form, and account for stochastic errors accurately. If these conditions are not met, specification errors occur, leading to biased or inefficient results. To address such problems, econometricians often rely on the Average Economic Regression (AER) or "bottom-up" approach, where models are gradually refined by adding variables based on diagnostics. Despite criticisms, this approach still holds relevance in econometric practice.

The selection of a model is guided by criteria such as parsimony, which emphasises simplicity without losing essential features; identifiability, which ensures unique parameter estimates; goodness of fit, measured by R^2; theoretical consistency, which requires coefficient signs to align with theory; and predictive power, which evaluates performance beyond the sample period.

Specification errors may arise from omitting relevant variables (underfitting), including unnecessary variables (overfitting), using an incorrect functional form, errors of measurement, or general model misspecification due to uncertainty about the true model. Each type of error has important consequences: underfitting leads to bias and inconsistency, while overfitting increases variances, reducing efficiency. Both undermine the reliability of hypothesis testing and inference.

Detection methods include t- and F-tests for assessing the significance of variables, residual analysis to identify patterns signalling omissions or functional form errors, the Durbin–Watson statistic to highlight misspecification, and Ramsey's RESET test, which introduces fitted value powers to detect errors through changes in R^2. Thus, correct specification is central to econometric modelling, requiring a careful balance between simplicity, accuracy, and theoretical soundness to minimise bias and inefficiency.

Assignments

1. Enumerate the main types of specification errors with empirical examples for each.
2. Define underfitting and overfitting within the bias-variance trade-off framework.
3. Explain the concept of model specification in the Classical Linear Regression Model (CLRM).
4. Evaluate the consequences of under-fitting (omitting a relevant variable) in a regression model.
5. Critically examine the methods used to detect specification errors in econometric models.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

BLOCK 3

Econometric Modelling

UNIT 1

The Three Variable Model

Learning Outcomes

After completing this unit, the learner will be able to:

- specify the population model and describe economic situations that require two regressors
- interpret OLS estimation of partial regression coefficients
- decompose variation for three-variable model

Background

A foundation in algebra, descriptive statistics, probability, and at least an introductory knowledge of simple linear regression is essential to study three variable models. These prerequisites ensure the student can understand not only the mechanics of multiple regression estimation but also the conceptual meaning behind regression coefficients and model fit measures.

Keywords

Multiple Regression, Three Variable Model, Coefficient of Determination (R^2), Adjusted R-Squared

Discussion

The concept of simple linear regression is applied where a single independent/predictor variable X was used to model the dependent/response variable Y. In most circumstances, there will be more than one independent variable that influences the response variable. Multiple regression models predict how a single response variable Y depends linearly on a number of predictor variables. Examples:

The selling price of a house can depend on the desirability of the location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage and a number of other factors.

The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

That is, we use the adjective "simple" to denote that our model has only one predictor, and we use the adjective "multiple" to indicate that our model has at least two predictors. The models have similar "LINE" assumptions. The only real difference is that whereas in simple linear regression we think of the distribution of errors at a fixed value of the single predictor, with multiple linear regressions we have to think of the distribution of errors at a fixed set of values for all the predictors.

A population model for a multiple linear regression model that relates a *y*-variable to *k* number of *x*-variables is written as,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Here we're using "*k*" for the number of predictor variables, which means we have *k*+1 regression parameters (the β coefficients).

We assume that the ϵi have a normal distribution with mean 0 and constant variance $\sigma 2$. These are the same assumptions that we used in simple regression with one *x*-variable.

The subscript *i* refers to the *i*th individual or unit in the population. In the notation for the *x*-variables, the subscript following *i(1,2...k)* simply denotes which *x*-variable it is.

The word "linear" in "multiple linear regression" refers to the fact that the model is *linear in the parameters*, $\beta_0, \beta_1, \ldots, \beta_k$. This simply means that each parameter multiplies an *x*-variable, while the regression function is a sum of these "parameter times *x*-variable" terms. Each *x*-variable can be a predictor variable or a transformation of predictor variables (such as the square of a predictor variable or two predictor variables multiplied together). Allowing non-linear transformation of predictor variables like this enables the multiple linear regression models to represent non-linear relationships between the response variable and the predictor variables.

The simplest form of multiple regression models is the three variable regression models.

3.1.1 Three Variable Regression Model

The simplest multiple regression model for two predictor variables is;

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \dots\dots\dots (1)$$

where β_0 is the intercept.

$\beta_1$measures the change in y with respect to x_1, holding other factors fixed.

β_2 measures the change in y with respect to x_2, holding other factors fixed.

Thus, a partial regression coefficient reflects the effects of one explanatory variable on the mean value of the dependent variable when the values of other explanatory variables included in the model are held constant.

In general, the three variable multiple regression model can be written as;

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \dots\dots\dots (2)$$

In the model with two independent variables, the key assumption about how u is related to x_1 and x_2 is;

$$E\ (u/x_1, x_2) = 0 \dots\dots\dots (3)$$

It means that, for any values of x_1 and x_2 in the population, the average of the unobserved factors is equal to zero. Given all other assumptions of classical model, it follows that, on taking the conditional expectation of y on both sides of the equation (2), we have;

$$E(y_i / x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \dots\dots\dots (4)$$

The equation gives, the conditional mean or expected value of y conditional upon the given or fixed values of the variables x_2 and x_3. Therefore, as in the two-variable case, multiple regression analysis is a regression analysis conditional upon the fixed values of the explanatory variables, and what we obtain is the average or mean value of y or mean response of y for the fixed values of x variables.

3.1.2 OLS Estimation of Partial Regression Coefficients

To estimate the parameters of the three variable regression model we consider the method of OLS. To find the OLS estimators, let us first write the sample regression function (SRF) corresponding to our PRF.

PRF: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$.....(2)

SRF: $Yi = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{u}_i$ .....(5)

From this we have,

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$.....(6)

where

$\hat{\beta}_0$ = the estimate of β_0.
$\hat{\beta}_1$ = the estimate of β_1.
$\hat{\beta}_2$ = the estimate of β_2.

But how do we obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? The method of **ordinary least squares** chooses the estimates to minimize the sum of squared residuals. That is, given n observations on y, $x1$, and $x2$, $\{(xi1, xi2, yi): i = 1, 2, \ldots, n\}$, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$

are chosen simultaneously to make the error sum of squares as small as possible. The error sum of squares is given as;

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \quad \text{.....(7)}$$

The most straight forward procedure to obtain the estimators that will minimise Residual Sum of Squares (RSS) is to differentiate it with respect to the unknowns and set the resulting expression equal to zero and finally solve them simultaneously. This procedure gives the normal equations as;

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2$$

$$\sum Y_i X_{i1} = \hat{\beta}_0 \sum X_{i1} + \hat{\beta}_1 \sum X_{i1}^2 + \hat{\beta}_2 \sum X_{i1} X_{i2}$$

$$\sum Y_i X_{i2} = \hat{\beta}_0 \sum X_{i2} + \hat{\beta}_1 \sum X_{i1} X_{i2} + \hat{\beta}_2 \sum X_{i2}^2$$

By simple algebraic manipulations of the preceding equations or simply by solving these normal equations, we obtain,

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2$$

$$\hat{\beta}_1 = \frac{(\sum y_i x_{i1})(\sum x_{i2}^2) - (\sum y_i x_{i2})(\sum x_{i1} x_{i2})}{(\sum x_{i1}^2)(\sum x_{i2}^2) - (\sum x_{i1} x_{i2})^2}$$

$$\hat{\beta}_2 = \frac{(\sum y_i x_{i2})(\sum x_{i1}^2) - (\sum y_i x_{i1})(\sum x_{i1} x_{i2})}{(\sum x_{i1}^2)(\sum x_{i2}^2) - (\sum x_{i1} x_{i2})^2}$$

Where;

- $y_i = Y_i - \bar{Y}$
- $x_{i1} = X_{i1} - \bar{X}_1$
- $x_{i2} = X_{i2} - \bar{X}_2$

3.1.3 Multiple Co-efficient of Determination R^2 and Adjusted R^2

Let R be the multiple correlation coefficient between y, and x_1, x_2,. x_k. Then square of multiple correlation coefficient (R^2) is called a coefficient of determination. The value of R^2 commonly describes how well the sample regression line fits the observed data. This is also treated as a measure of **goodness of fit** of the model.

Assuming that the intercept term is present in the model as

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + u_i,\ i = 1, 2, \ldots, n$$

Then,

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= 1 - \frac{SS_{res}}{SS_T} = \frac{SS_{reg}}{SS_T}$$

Where SS_{res} = sum of squares due to residuals,

SS_T = total sum of squares

$SS\ reg$ = sum of squares due to regression.

R^2 measures the explanatory power of the model, which in turn reflects the goodness of fit of the model. It reflects the model adequacy in the sense of how much is the explanatory power of the explanatory variables.

The limits of R^2 are 0 and 1, i.e.,

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

$R^2 = 0$ indicates the poorest fit of the model.

$R^2 = 1$ indicates the best fit of the model

For example, $R^2 = 0.95$ indicates that 95% of the variation in y is explained by R^2. In simple words, the model is 95% good. Similarly, any other value of R^2 between 0 and 1 indicates the adequacy of the fitted model.

Adjusted R^2

If more explanatory variables are added to the model, then R^2 increases. In case the variables are irrelevant, then R^2 will still increase and gives an overly optimistic picture. With a purpose of correction in the overly optimistic picture, adjusted R^2, denoted as $\bar{R}^2$ or adj R^2 is used which is defined as,

$$\bar{R}^2 = 1 - \frac{SS_{res}/(n-k)}{SS_T/(n-1)}$$
$$= 1 - \left(\frac{n-1}{n-k}\right)(1-R^2)$$

We will see later that $(n - k)$ and $(n - 1)$ are the degrees of freedom associated with the distributions of SS_{res} and SS_T.

Moreover, the quantities $\frac{SS_{res}}{n-k}$ and $\frac{SS_T}{n-1}$ are based on the unbiased estimators of respective variances of e and y in the context of analysis of variance.

The adjusted R^2 will decline if the addition if an extra variable produces too small a reduction in $(1- R^2)$ to compensate for the increase in $\left(\frac{n-1}{n-k}\right)$. Another limitation of adjusted R^2 is that it can be negative also. For example, if $k = 3$, $n = 10$, $R2 = 0.16$, then

$\bar{R}^2 = 1 - \frac{9}{7} \times 0.97 = -0.25 < 0$ which has no interpretation.

Limitations of R^2

If the constant term is absent in the model, then R^2 cannot be defined. In such cases, R^2 can be negative. Some ad- hoc measures based on R^2 for regression line through origin have been proposed in the literature.

R^2 is sensitive to extreme values, so R^2 lacks robustness.

R^2 always increases with an increase in the number of explanatory variables in the model. The main drawback of this property is that even when the irrelevant explanatory variables are added in the model, R^2 still increases. This indicates that the model is getting better, which is not really correct.

Summarised Overview

Multiple regression extends the concept of simple linear regression, which involves a single predictor variable, to situations where two or more independent variables jointly explain the variation in a dependent variable. For example, the selling price of a house may depend on location, size, number of rooms, and year built, while a child's height may depend on parental height, nutrition, and environment. The term "simple" denotes one predictor, while "multiple" denotes at least two, but both share the same classical regression assumptions. A general population model for multiple regression is expressed as $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$, where k is the number of predictors and ϵ_i are normally distributed errors with mean zero and constant variance. The model is called "linear"

because parameters enter linearly, even though predictor variables can be transformed (squared terms, interactions) to capture non-linear effects.

The simplest case is the three-variable regression model: $y=\beta_0+\beta_1x_1+\beta_2x_2+u$. Here, β_1 and β_2 are partial regression coefficients measuring the effect of each variable on y while holding the other constant. In estimation, the Ordinary Least Squares (OLS) method is applied to choose parameter estimates that minimise the Residual Sum of Squares (RSS). This leads to normal equations which, once solved, provide unbiased and consistent estimates of the parameters.

The explanatory power of a regression model is summarised by the coefficient of determination, R^2, which measures the proportion of variation in the dependent variable explained by the model. Its value ranges between 0 and 1, with values closer to 1 indicating a better fit. However, since R^2always increases with the addition of predictors, even irrelevant ones, adjusted $R^2$is used to account for the number of variables and sample size. Adjusted R^2may decline if an additional predictor does not sufficiently improve explanatory power, and in some cases, it may even take negative values, making interpretation problematic. Moreover, $R^2$is sensitive to extreme values and cannot be defined if the intercept term is omitted. Despite such limitations, multiple regression remains a powerful tool for quantifying relationships when multiple factors influence an outcome.

Assignments

1. Decompose variation TSS = ESS +RSS and define R^2 = ESS/TSS for the three-variable model.
2. State the population three-variable linear model and define each symbol clearly.
3. In multiple regression, what does “partial” in “partial regression coefficient” mean? Illustrate your answer when both x_1 and x_2 are in the model.
4. Explain why R^2 never falls when an additional regressor enters and why that can be misleading for model comparison.

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 2

Dummy Variables and ANOVA Models

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the concept of dummy variable and how dummy variable regression can be used for deriving inferences
- estimate and compare ANOVA and ANCOVA models
- learn to avoid dummy variable trap

Background

Economic contexts often incorporate qualitative variables as well in explaining cause-effect relationships. Hence how categorical data can be coded and analysed, which is foundational for dummy variable construction should be looked into.

When you introduce dummy variables, you're essentially using regression to code group membership (categorical predictors like "treatment" vs "control," or "male" vs "female") so you can measure their effect within a regression framework. Previous regression training helps you see that the coefficient for a dummy variable represents the average difference in the outcome between the reference and coded group, controlling for other factors.

A solid foundation in basic statistics, regression analysis, and categorical data handling is required for a student to confidently master these intermediate and advanced quantitative methods.

Keywords

Qualitative Explanatory Variables, Dummy Variable, Dummy Variable Regression, Dummy Variable Trap, Multicollinearity, ANOVA (Analysis of Variance), ANCOVA (Analysis of Covariance)

Discussion

3.2.1 Qualitative Explanatory Variables

In regression analysis, we often deal with quantitative variables such as income, output, prices, or age. However, many important determinants of economic and social behaviour are qualitative in nature. These include attributes like gender, race, religion, region, marital status, or political affiliation. Although these variables are not numerical in the usual sense, they can still be incorporated into regression models through a special technique: the use of dummy variables (also called indicator variables or categorical variables).

Qualitative variables represent categories rather than magnitudes. For instance:

- Gender: male or female
- Region: South vs. non-South
- Education: graduate vs. non-graduate

Such variables may influence the dependent variable (for example, wages, consumption, or voting behaviour). Ignoring them can lead to biased or incomplete models, as the effects of such attributes would be wrongly absorbed into the error term.

3.2.2 Dummy Variable

A dummy variable is an artificial variable used in regression analysis to represent qualitative or categorical factors such as gender, region, religion, or marital status. Since regression models require numerical inputs, such categories are converted into binary codes, usually taking the value 1 if the attribute is present and 0 if absent. For example, in studying wage differences, a variable coded as 1 for female and 0 for male allows gender to enter the regression model. The coefficient of the dummy variable then measures the difference in the dependent variable (e.g., wages) between the category coded as 1 and the base category coded as 0. One category is always left out as the reference group to avoid the dummy variable trap, which is the problem of perfect multicollinearity when all categories are included simultaneously. Dummy variables are widely applied in economics to study group differences, seasonal effects, and policy impacts, and they can also be interacted with quantitative variables to capture differences in relationships across groups. In short, dummy variables provide a systematic way of incorporating qualitative information into quantitative regression models.

3.2.3 Use of Dummy Variables

Dummy variables are widely used in econometrics because they allow us to incorporate qualitative information into regression models. Their main uses include:

1. **Measuring Group Differences**: They help capture differences between categories such as gender (male/female), marital status (married/unmarried), caste, or region (urban/rural, north/south). For example, they can show whether female workers earn less than male workers, holding other factors constant.

2. **Policy Impact Analysis**: Dummy variables are useful in evaluating the effect of policies or events by assigning 0 to the period before implementation and 1 to the period after. This helps to quantify the impact of policy changes, reforms, or economic shocks.

3. **Seasonal and Time Effects**: In time-series analysis, dummies can account for seasonality (quarters, months, festivals) or specific years (crisis years, pandemic periods), thereby isolating recurring patterns.

4. **Structural Breaks and Chow Test**: They are used to test whether relationships between variables change over time or across groups. For example, we can check if demand functions differ before and after liberalisation.

5. **Interaction Effects**: By combining dummy variables with quantitative variables, researchers can study whether slopes differ across groups, such as whether education increases wages differently for men and women.

6. **Piecewise or Switching Regressions**: Dummies can divide the sample into different regimes, allowing separate intercepts or slopes in different ranges of data, such as before and after a threshold income level.

3.2.4 Regression on Dummy Variables

In regression analysis the dependent variable is frequently influenced not only by variables that can be readily quantifiable but also by variables that are qualitative in nature like sex, race, colour, religion, nationality etc. For example, holding all other factors constant, female workers are founded to be earning less than their male counterparts and non-whites are found to earn less than whites. This pattern may result from sex or racial discrimination, but whatever the reason qualitative variables such as sex and race do influence the dependent variable and clearly should be included among the explanatory variables.

Since qualitative variables are usually indicate the presence or absence of a quality or an attribute, such as male or female black or white, one method of quantifying such attributes is by constructing artificial variables that taken on values 1 or 0, 0 indicating the absence of an attribute and 1 indicating the presence of that attribute. For example, 1 may indicate that a person is a male and 0 may designate a female. Or 1 may indicate a person is a graduate and 0 that he is not and so on.

Variables that assume such 0 and 1 values are called dummy variables. Alternative names are:

- Indicator variables
- Binary variables
- Categorical variables
- Qualitative variables
- Dichotomous variables

3.2.5 ANOVA Models

Dummy variables can be used in regression models just as easily as quantitative variables. Here regression model contains explanatory variables that are exclusively dummy variables are called Analysis of Variance (ANOVA) models. For example,

$Y_i = \alpha + \beta D_i + u_i$ .....(1)

Where;

Y_i = salary of a worker

D_i = 1, if male

= 0, if female

In (1), instead of quantitative X variable we have a dummy variable D. Model (1) enable us to find out whether sex make any difference in the salary of a worker, if all other variables such as age, education, years of experience etc. are held constant. If u_i satisfies all the assumptions of CLRM, We obtain from (1),

Mean salary of a female worker,

$E(Y_i / D_i = 0) = \alpha$ ..... (2)

Means salary of a male worker,

$E(Y_i / D_i = 1) = \alpha + \beta$ ..... (3)

That is, the intercepted term 'α' gives the mean salary of a female worker and the slope coefficient 'β' tells by how much the means salary of a male worker differs from the means salary of his female counterparts, α +β reflecting the mean salary of a male worker.

A test of the H_0: β= 0, that is, there is no sex discrimination can be easily made by running regression on (1) in usual manner and finding out whether on the basis of the 't' test the estimated β is statistically significant.

3.2.6 ANCOVA Models

In most economic research, a regression model contains some explanatory variables that are quantitative and some that are qualitative. Regression models containing a mixture of quantitative and qualitative variables are called Analysis of Co-Variance (ANCOVA) models.

An ANCOVA model is,

$Y_i = \alpha_1 + \alpha_2 D_i + \beta X_i + u_i$ .....(4)

Where; Y_i = salary of a worker

X_i = Years of experience

D_i = 1, if male

= 0, if female

Model (4) contains one quantitative variable (years of experience) and one qualitative variable (sex) that has two classes or categories namely male and female. (4) means that,

Mean salary of a female worker,

$$E(Y_i / X_i, D_i = 0) = \alpha + \beta X_i \quad \text{.....(5)}$$

Mean salary of a male worker,

$$E(Y_i / X_i, D_i = 1) = \alpha_1 + \alpha_2 + \beta X_i \quad \text{.....(6)}$$

Model (4) postulates that the male and female workers salary functions in relation to the years of experience have the same slope (β) but a different intercept. In other words, it is assumed that the level of male workers' salary is different from the means salary of female workers by α_2 but the rate of change in the mean salary by years of experience is same for both sexes.

If the assumption of common slope is valid, a test of the hypothesis that the two regressions (5) and (6) have the same intercept, (that is there is no sex discrimination) can be made easily by running the regression on (4) and noting the statistical significance of the estimated α_2 on the basis of 't' test. If the 't' test shows that α_2 is statistically significant, we reject the null hypothesis that the male and female workers' level of mean salary are the same.

3.2.7 Interpretation of Regression Results

There are several methods of de-seasonalising a time series and the method of dummy variables is one of the popular methods. For this, we are using the regression equation as follows

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} \beta X_i + u_i \quad \text{.....(8)}$$

Where;

Y_i = Profits X_i = Sales

D_{2i} = 1, if Second quarter

= 0, If otherwise

D_{3i} = 1, Third quarter

= 0, if otherwise

D_{4i} = 1, First quarter

= 0, if otherwise

u_i = Stochastic error term

For this, firstly we have quarter-wise data and we assign values for each quarters using dummy variables. Note that we are assuming that, the variable 'season' has four classes, the four quarters of a year, thereby requiring the use of three dummy variables. Thus, if there is a seasonal pattern present in various quarters and if it is statistically significant, the estimated differential intercepts α_2, α_3, and α_4, will reflect it. It is possible that only some of these differential intercepts are statistically significant so that only some quarters may reflect it. The above model is general enough to accommodate all these cases.

The required number of dummies in a dummy variable regression model is always one less than the number of categories for any qualitative variable included in the model.

Rule for Dummy Variable Construction

If a qualitative variable has m categories, the regression model must include m−1

m−1 dummy variables for that variable.

This avoids the problem known as the dummy variable trap (perfect multicollinearity), which occurs if all categories are coded as dummies together with an intercept term.

In this model it was assumed that only the intercept term differs between quarters, the slope coefficient of the sales variable being the same in each quarter.

We can use dummy variables in another case called piecewise linear regression analysis. This case occurs when trend line occurs with different slopes. Suppose consider a case of a company remunerates its sales representatives. It pays commission based on sales in such a manner that up to a certain level the target threshold (level X*) there is one commission structure and beyond that level another. it can be depicted in the figure below.

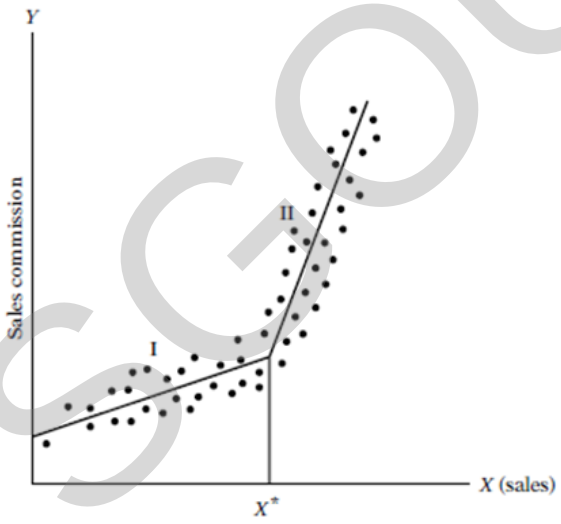


Fig.3.2.1 Piecewise Linear Regression Analysis

More specifically, it is assumed that commission for sales increases linearly with an increase in sales until the threshold level X*, after which also it increases linearly with sales but at a much steeper rate. Thus, we have a piece-wise linear regression consisting of two linear pieces or segments, which are labelled I and II in the Figure above and the commission function changes its slope at the threshold value. Given the data on commission, sales and the value of the threshold level X*, the technique of the many variables can be used to estimate the differing slopes of the two segments of the piece- wise linear regression shown in Figure above. Thus, we proceed as,

$$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i \quad \text{.....(9)}$$

Where;

Y_i =Sales Commission X_i =Volume of sales

X^* = Threshold value of sales *(not known in advance)*

$D_i = 1$ if $X_i > X^*$

$= 0$ if $X_i < X^*$ Assuming $E(u_i) = 0$,

$E(Y_i / D_{i=1}, X_i, X^*) = \alpha_1 - \beta_2 X^* + (\beta_1 + \beta_2) X_i$ which gives the mean sales commission beyond the target level X^* and $E(Y_i / D_{i=0}, X_i, X^*) = \alpha_1 + \beta_1 X_i$ gives the mean sales commission up to the target level X^*.

Thus, β_1 gives the slope of the regression line in segment I and $(\beta_1+\beta_2)$ gives the slope of the regression line segment II of the piece-wise linear regression shown in Figure above. A test of the hypothesis that there is no break in the regression at the threshold value X^* can be conducted easily by noting the statistical significance of the estimator differential slope coefficient β_2 as in the Figure below.

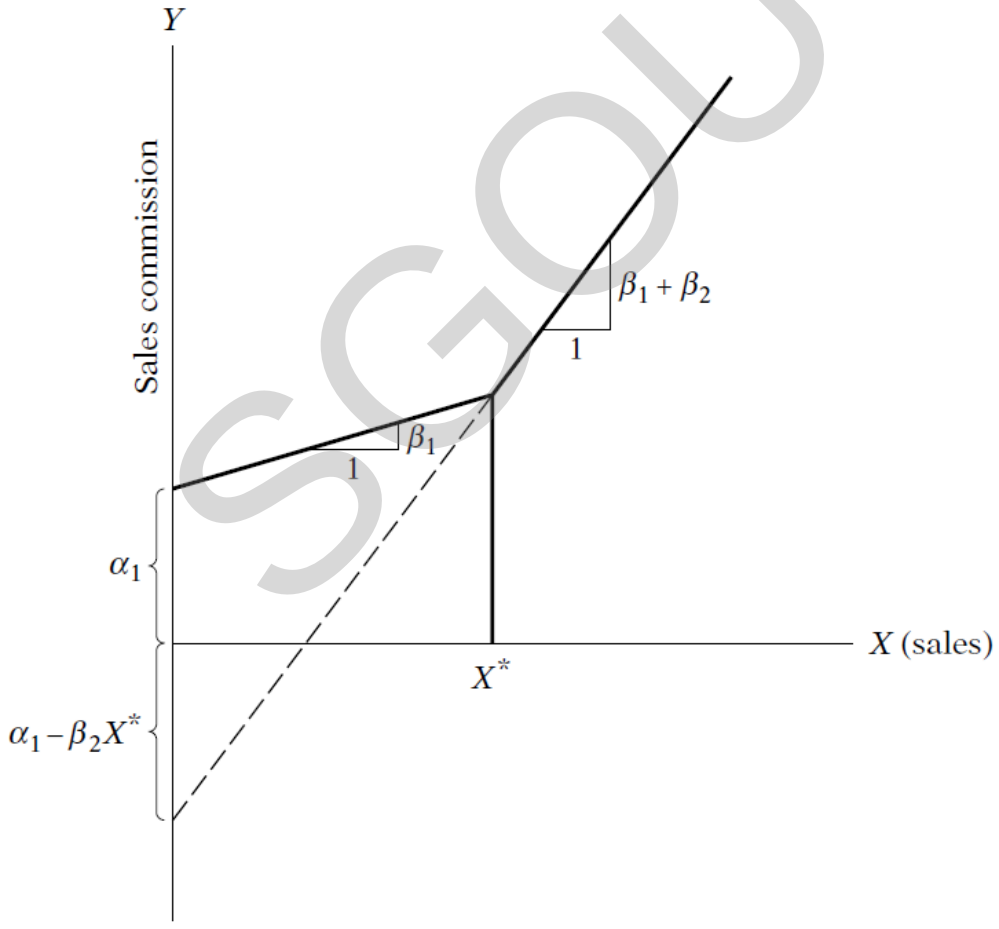


Fig 3.2.2 Regression Line

3.2.8 Dummy Variable Trap

To distinguish the two categories of a dummy variable, we have introduced only one dummy variable D_i. If, D_i=1, always denote a male, when D_i=0 we know that it is a female since there are only two possible outcomes. Hence one dummy variable suffices to distinguish two categories. Let us the model is as,

$$Y_i = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + u_i \quad \ldots\ldots\ldots\ldots (7)$$

Where;

Y_i = salary of a worker

X_i = Years of experience

D_{2i} = 1, if male

= 0, if female

D_{3i} = 1, if female

= 0, if male

The model (7) cannot be estimated because of perfect collinearity between D_2 and D_3. To see this, suppose we have a sample of three male workers and two female workers as follows,

Table 3.2.1 Sample Data

	Y	α_1	D_2	D_3	X
Male	Y_1	1	1	0	X_1
Male	Y_2	1	1	0	X_2
Female	Y_3	1	0	1	X_3
Male	Y_4	1	1	0	X_4
Female	Y_5	1	0	1	X_5

It is clear from the data that, $D_2 = 1- D_3$ or $D_3 = 1- D_2$

That is, D_2 and D_3 are perfectly collinear. In case of perfect multi-collinearity, it is clear that the usual OLS estimation is not possible. One simple way to avoid this problem is that to assign only one dummy variable if there are only two levels or classes of the qualitative variable. Thus, the general rule is that if a qualitative variable has 'm' categories, introduce 'm-1' dummy variables. If this rule is not followed, we shall fall into what might be called the dummy variable trap. That is the situation of perfect multi-collinearity.

The dummy variable trap is a scenario in which the independent variables are multicollinear– a scenario in which two or more variables are highly correlated; in simple terms one variable can be predicted from the others. In short, a dummy variable model with perfect or high multi-collinearity is the situation of 'dummy variable trap'.

Summarised Overview

In regression analysis, explanatory variables are not always quantitative; many important determinants of economic and social behaviour are qualitative in nature, such as gender, race, region, or marital status. These categorical variables can be incorporated into regression models through the use of dummy variables, also called indicator or binary variables. Dummy variables take values of 0 and 1 to indicate the absence or presence of a particular attribute, allowing non-numeric factors to be quantified and analysed. For instance, in studying wage differences, coding 1 for female and 0 for male permits the regression model to capture the effect of gender on wages. The coefficient of such a dummy variable measures the difference in the dependent variable between the coded category and the base category, while one category is always omitted to avoid perfect multicollinearity, known as the dummy variable trap. Dummy variables are widely used in econometrics to study group differences, measure policy impacts, capture seasonal effects, account for structural breaks, model interaction effects, and represent piecewise regressions.

Regression models that contain only dummy variables are referred to as ANOVA models, while those containing both quantitative and qualitative variables are called ANCOVA models. ANOVA models test for differences in mean values across categories—for example, male versus female salaries—by estimating separate intercepts. ANCOVA models, on the other hand, combine continuous explanatory variables with dummy variables, such as examining how salaries vary by both years of experience and gender, allowing for different intercepts across groups while maintaining common slopes. Dummy variables can also be employed in time-series contexts to capture seasonal patterns, where quarterly or monthly effects are isolated by introducing multiple dummy variables, ensuring that one category is excluded as a base. Similarly, in piecewise linear regression, dummy variables can capture structural changes at threshold values, such as commission rates that change once a sales target is reached.

However, caution is necessary to avoid the dummy variable trap, which arises when all categories of a qualitative variable are coded into the model, resulting in perfect multicollinearity. For instance, if male and female are both introduced as separate dummy variables along with an intercept, one can be expressed as a linear function of the other, making OLS estimation impossible. The general rule is that if a qualitative variable has m categories, only $m–1$ dummy variables should be included. By carefully constructing and interpreting dummy variables, econometricians can systematically incorporate qualitative information into regression models, thereby enriching analysis and improving model accuracy.

Assignments

1. Show that a pure-dummy regression with no quantitative covariates reproduces the one-way ANOVA model $y_{ij} = \mu + \tau_j + u_{ij}$
2. Derive group means from estimated coefficients and test the null hypothesis of equal means via an overall F-test.
3. Explain how ANCOVA controls for covariates while comparing adjusted group means or slopes.
4. Define the dummy variable trap as perfect multicollinearity caused by including all category dummies plus an intercept. How can you avoid dummy variable trap?
5. Define a qualitative (categorical) explanatory variable and describe how it is represented in a linear regression.
6. Explain the difference between a one-way ANOVA model and a regression that contains only a set of group dummies plus an intercept.
7. Express the ANOVA model in regression form with dummy variables.
8. When an interaction between a continuous variable and a dummy is added to a regression, what is the economic meaning of the interaction coefficient?

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 3

Qualitative Response Models

Learning Outcomes

After completing this unit, the learner will be able to:

- equip to choose, estimate, diagnose, and interpret qualitative response models in applied econometric research
- use LPM as a baseline or quick exploratory tool before moving to nonlinear models
- helps to use maximum likelihood (ML) or limited-information methods to estimate nonlinear models and assess convergence diagnostics
- interpret estimated coefficients in terms of probabilities, marginal effects, elasticities, and latent variables

Background

This unit introduces students to a class of econometric models designed to handle qualitative or categorical dependent variables, commonly referred to as qualitative response models. Unlike classical linear regression models—which are suited for continuous outcomes—these models allow researchers to rigorously analyze scenarios where the outcome variable is binary, limited, or censored, a frequent requirement in economics and the social sciences.

Keywords

Qualitative Response Models, Linear Probability Model (LPM), Logit Model, Probit Model, Tobit Model

Discussion

3.3.1 Qualitative Response Models

Qualitative Response models are models in which the regressand itself is qualitative in nature. Although increasingly used in various areas of social sciences and medical research, qualitative response regression models pose interesting estimation and interpretation challenges. In this unit we are discussing some of the major themes in this area.

3.3.2 LPM (Linear Probability Model)

LPM refers to Linear Probability Model. To fix ideas, consider the following regression model:
$Y_i = \beta_1 + \beta_2 X_i + u_i$ ---------- (1)

Where,

X = family income and

Y =1 if the family owns a house and 0 if it does not own a house.

The above model looks like a typical linear regression model but because the regressand is binary, or dichotomous, it is called a Linear Probability Model (LPM). This is because the conditional expectation of Y_i given Xi, $E(Y_i/X_i)$, can be interpreted as the conditional probability that the event will occur given Xi that is, Pr $(Y_i = 1/X_i)$ Thus, in our example, $E(Y_i/X_i)$ Gives the probability of a family owning a house and whose income is the given amount X_i.

The justification of the name LPM for model like (1) can be seen as follows: Assuming $E(u_i) = 0$ usual (to obtain unbiased estimators) we obtain.

$$E(Y_i/X_i) = \beta_1 + \beta_2 X_i \quad (2)$$

Now, if P_i = probability that $Y_i = 1$ (that is, the event occurs), and $(1-P_i)$ = probability that $Y_i = 0$ (that is, that the event does not occur), the variable Y_i has the following (probability) distribution.

Table 3.3.1 Probability Distribution

Y_i	Probability
0	$1-P_i$
1	P_i
Total	1

That is Y_i follows the Beronoulli probability distribution. Now, by the definition of mathematical expectation, we obtain.

$$E(Y_i) = 0(1-P_i) + 1\ (P_i) = P_i \text{ .....(3)}$$

Comparing (2) with (3.), we can equate, $E(Y_i/X_i) = \beta_1 + \beta_2 X_i = P_i$ .....(4)

That is, the conditional expectation of the model can, in fact, be interpreted as the conditional probability of Y_i in general; the expectation of a Bernoulli random variable is the probability that the random variable equals 1. In passing note that if there are 'n 'independent trials, each with a probability p of success and probability (1-p) of failure, and X of these trials represent the number of successes then X is said to follow the binomial distribution. The mean of the binomial distribution is 'np' and its variance is 'm(1-p)'. The term success is defined in the context of the problem.

Since the probability P_i must lie between 0 and 1, we have the restriction.

$$0 \leq E(Y_i/X_i) \leq 1 \ldots\ldots\ldots\ldots 5$$

That is, the conditional expectation (or conditional probability) must lie between 0 and 1.

From the preceding discussion it would seem that OLS can be easily extended to binary dependent variable regression models. So, perhaps there is nothing new here. Unfortunately, this is not the case, for the LPM poses several problems, which are as follows:

Non-Normality of the Disturbances u_i.

Although OLS does not require the disturbances (u_i) to be normally distributed, we assumed them to be so distributed for the purpose of statistical inference. But the assumption of normality for u_i is not tenable for the LPMs because, like Y_i, the disturbances u_i also take only two values; that is, they also follow the Bernoulli distribution. This can be seen clearly if we write equation (1) as,

$$u_i = Y_i - \beta_1 - \beta_2 X_i \ldots\ldots\ldots\ldots (6)$$

The probability distribution of u_i is,

Table 3.3.2 Probability Distribution

	U_i	Probability
When Y_i=1	$1 - \beta_1 - \beta_2$	Xi Pi
When Y_i=0	$- \beta_1 - \beta_2 X_i$	1-Pi

..... (7)

Obviously, u_i cannot be assumed to be normally distributed; they follow the Bernoulli distribution. But the non-fulfilment of the normality assumption may not be as critical as it appears because we know that the OLS point estimates still remain unbiased (recall that, if the objective is point estimation, the normality assumption is not necessary). Besides, as the sample size increases indefinitely, statistical theory shows that the OLS estimators tend to be normally distributed generally. As a result, in large samples the statistical inference of the LPM will follow the usual OLS procedure under the normality assumption.

Heteroscedastic Variances of the Disturbances:

Even if $E(u_i) = 0$ and cov $(u_i , u_j) = 0$ for $i \_= j$ (i.e., no serial correlation), it can no longer be maintained that in the LPM the disturbances are homoscedastic. This is, however, not surprising. As statistical theory shows, for a Bernoulli distribution the theoretical mean and variance are, respectively, p and $p(1 - p)$, where p is the probability of success (i.e., something happening), showing that the variance is a function of the mean. Hence the error variance is heteroscedastic.

For the distribution of the error term given in (7), applying the definition of variance, the reader should verify that,

$$\text{var } (u_i) = P_i(1 - P_i) \quad \text{.....(8)}$$

That is, the variance of the error term in the LPM is heteroscedastic. Since $Pi = E(Y_i \mid X_i) = \beta_1 + \beta_2 X_i$, the variance of u_i ultimately depends on the values of X and hence is not homoscedastic.

We already know that, in the presence of heteroscedasticity, the OLS estimators, although unbiased, are not efficient; that is, they do not have minimum variance. But the problem of heteroscedasticity, like the problem of non-normality, is not insurmountable. Since the variance of u_i depends on $E(Y_i|X_i)$, one way to resolve the heteroscedasticity problem is to transform the model (1) by dividing it through by,

$$\sqrt{E(Y_i/X_i)}\,[1 - E(Y_i/X_i)] = \sqrt{P_i}\,(1 - P_i) = \text{say } \sqrt{w_i}$$

That is,

$$Y_i / \sqrt{w_i} = \sqrt{\beta_1} / w_i + \beta_2 X_i / \sqrt{w_i} + u_i / \sqrt{w_i} \text{ ..... (9)}$$

As you can readily verify, the transformed error term in (9) is homoscedastic. Therefore, after estimating (1), we can now estimate (9) by OLS, which is nothing but the *Weighted Least Squares* (WLS) with w_i serving as the weights. In theory, what we have just described is fine. But in practice the true $E(Y_i \mid X_i)$ is unknown; hence the weights w_i are unknown. To estimate w_i, we can use the following two-step procedure:

Step 1. Run the OLS regression (1) despite the heteroscedasticity problem and obtain $\hat{Y}_i$ = estimate of the true $E(Y_i \mid X_i)$. Then obtain

$$\hat{w}_i = \hat{Y}_i (1 - \hat{Y}_i), \text{ the estimate of } w_i.$$

Step 2. Use the estimated w_i to transform the data as shown in

(9) and estimate the transformed equation by OLS (i.e., weighted least squares).

Non-fulfillment of $0 \leq E(Yi \mid X) \leq 1$

Since $E(Y_i \mid X)$ in the linear probability models measures the conditional probability of the event Y occurring given X, it must necessarily lies in between 0 and 1. Although this is true a priori, there is no guarantee that $\hat{Y}_i$, the estimators of $E(Y_i \mid X_i)$, will necessarily fulfil this restriction, and this is the real problem with the OLS estimation of the LPM. There are two ways of finding out whether the estimated $\hat{Y}_i$ lie between 0 an1. One is to estimate the LPM by

the usual OLS method and find out whether the estimated $\hat{Y}_i$ lie between 0 and 1. If some are less than 0 (that is, negative), $\hat{Y}_i$ is assumed to be zero for those cases; if they are greater than 1, they are assumed to be 1.

The second procedure is to devise an estimating technique that will guarantee that the estimated conditional probabilities $\hat{Y}_i$ will lie between 0 and 1. The logit and probit models discussed later will guarantee that the estimated probabilities will indeed lie between the logical limits 0 and 1.

3.3.3 Logit and Probit Models

The LPM is plagued by several problems, such as (1) non-normality of u_i, (2) heteroscedasticity of u_i, (3) possibility of $\hat{Y}_i$ lying outside the 0–1 range, and (4) the generally lower $R2$ values. But these problems are surmountable. For example, we can use WLS to resolve the heteroscedasticity problem or increase the sample size to minimize the non-normality problem. By resorting to restricted least-squares or mathematical programming techniques we can even make the estimated probabilities lie in the 0–1 interval. But even then, the fundamental problem with the LPM is that it is not logically a very attractive model because it assumes that,

$P_i = E\ (Y = 1/X)$ increases linearly with X, that is, the marginal or incremental effect of X remains constant throughout. This seems patently unrealistic. In reality one would expect that P_i is nonlinearly related to X_i:

At very low income a family will not own a house but at a sufficiently high level of income, say, X^*, it most likely will own a house. Any increase in income beyond X^* will have little effect on the probability of owning a house. Thus, at both ends of the income distribution, the probability of owning a house will be virtually unaffected by a small increase in X. Therefore, what we need is a (probability) model that has these two features:

1. As Xi increases, $Pi = E\ (Y = 1 \mid X)$ increases but never steps outside the 0–1 interval, and

2. the relationship between Pi and Xi is nonlinear, that is,

—one which approaches zero at slower and slower rates as Xi gets small and approaches one at slower and slower rates as Xi gets very large.

The reader will realize that the sigmoid, or S-shaped, curve very much resembles the cumulative distribution function (CDF) of a random variable. Therefore, one can easily use the CDF to model regressions where the response variable is dichotomous, taking 0–1 values. The practical question now is, which CDF? For although all CDFs are S shaped, for each random variable there is a unique CDF. For historical as well as practical reasons, the CDFs commonly chosen to represent the 0–1 response models are

1. The logistic and
2. The normal,

The former giving rise to the **logit** model and the latter to the **probit** (or **normit**) model.

3.3.3.1 Logit Model

$P_i = E(Y = 1/X_i) = \beta_1 + \beta_2 X_i$ 1

Where X is income and Y = 1 means the family owns a house. But now consider the following representation of home ownership:

$$P_i = E(Y = 1/X_i) = \frac{1}{1+e^{-(\beta_1+\beta_2 X_i)}} \text{ ----------2}$$

For ease of exposition, we write (2) as

$$P_i = \frac{1}{1+e^{-Z_1}} = \frac{e^z}{1+e^z} \text{ ----------3}$$

Where $Z_i = \beta_1 + \beta_2 X_i$

Equation (3) represent what is known as the cumulative logistic distribution function.

It is easy to verify that as Z_i ranges from - ∞ to + ∞, P_i Ranges between 0 and 1 and that P_i is nonlinearly related to Z_i (i.e.X_i). Thus satisfying the two requirements considered earlier. But it seems that in satisfying these requirements, we have created an estimation problem because P_i is nonlinear not only in X but also in the β's as can be seen clearly from (2). This means that we cannot use the familiar OLS procedure to estimate the parameters. But this problem is more apparent than real because (2) can be linearised, which can be shown as follows.

If P_i the probability for owning a house, is given by (3) then (1-Pi), the probability of not owning a house, is

$$1 - P_i = \frac{1}{1+e^{-Z_1}} \text{ ----------4}$$

Therefore, we can write,

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i} \text{ ----------5}$$

Now $P_i / (1 - P_i)$ is simply the odds ratio in favour of owning a house, the ratio of the probability that a family will own a house to the probability that it will not own a house. Thus if P_i

= 0.8, it means that odds are 4 to 1 in favour of the family owning a house.

Now if we take the natural log of (5) we obtain a very interesting result, namely,

$$L_i = In \frac{P_i}{1-P_i} = Z_i \text{ .....(6)}$$

$= \beta_1 + \beta_2 X_i$

That is, L, the log of the odds ratio, is not only linear in X, but also (from the estimation viewpoint) linear in the parameters. L is called the logit, and hence the name logit model for models like (6) Notice these features of the logit model.

1. As P goes from 0 to 1 (i.e. as Z varies from - ∞ to + ∞), the logit L goes from - ∞ to + ∞. That is, although the probabilities (of necessity) lie between 0 and 1, the logits are not so bounded.

2. Although l is linear in X, the probabilities themselves are not. This property is in contrast with the LPM model (1) where the probabilities increase linearly with X.

3. Although we have included only a single X variable, or regressor, in the preceding model, one can add as many regressors as may be dictated by the underlying theory.

4. If L, the logit, is positive, it means that when the value of the regressor(s) increases, the odds that the regressand equals 1 (meaning some event of interest happens) increases. If L is negative, the odds that the regressand equal 1 decreases as the value of X increases. To put it differently, the logit becomes negative and increasingly large in magnitude as the odds ratio decreases from 1 to 0 and becomes increasingly large and positive as the odds ratio increases from 1 to infinity.

5. More formally, the interpretation of the logit model given in (6) is as follows: β_2, the slope, measures the change in L for unit change in X, that is, it tells how the log – odds in favor of owning a house change an income changes by a unit, say $ 1000. The intercept β_1 is the value of the log odds in favour of owning a house if income is zero. Like most interpretations of intercepts, this interpretation may not have any physical meaning.

Given a certain level of income, say, X, if we actually want to estimate not the odds in favor of owning a house but the probability of owning a house itself, this can be done directly from (3) once the estimate of $\beta_1 + \beta_2$ are available. This, however, raises the most important question. How do we estimate β_1 and β_2 in the first place? The answer is given in the next section.

Whereas the LPM assumes that Pi is linearly related to X_i the logit model assumes that the log of the odds ratio is linearly related to X_i.

3.3.3.2 Probit Model

The estimating model that emerges from the normal CDF is popularly normit model. To motivate the probit model, assume that in our home ownership example the decision of the i^{th} Family to own a house or not depends on an unobservable utility index I_i (also known as a latent variable), that is determined by one or more explanatory variables, say income X_i in such a way that the larger the value of the index I_i The greater the probability of a family owning a house. We express the index I_i as.

$$I_i = \beta_1 + \beta_2 X_i \ldots\ldots\ldots\ldots 1$$

Where; X_i is the income of the i^{th} Family.

How is the (unobservable) index related to the actual decision to own a house? As before let Y = 1if the family owns a house and Y = 0 if it does not. Now it is reasonable to assume that there is a critical or threshold level of the index, call it I^*_i, such that if I_i exceeds I^*_i, the family will own a house, otherwise it will not. The threshold I^*_i like I_i is not observable, but if we assume that it

is normally distributed with the same mean and variance, it is possible not only to estimate the parameters of the index given in (1). But also to get some information about the unobservable index itself. This calculation is also follows.

Given the assumption of normality, the probability that I^*_i is less than or equal to I_i can be computed from the standardized normal CDF as.

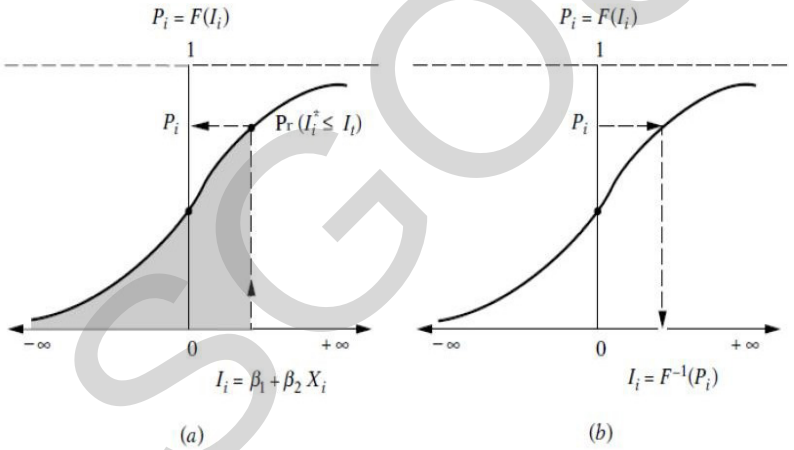
$P_i = P(Y=1/X) = P(I_i^* \leq I_i) = P(Z_i \leq \beta_1 + \beta_2 X_i) = F(\beta_1 + \beta_2 X_i)$ ----------2

Where P(Y = 1/X) means the probability that an event occurs given the value(s) of the X, or explanatory, variable(s) and where Zi is the standard normal variable, i.e. $Z \sim N(0\text{-}\sigma^2)$. F is the standard normal CDF, which written explicitly in the present context is:

$$F(I_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{I_i} e^{-z^{2/2}} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_i + \beta_2 X_i} e^{-z^{2/2}} dz \quad \text{.....(3)}$$

Since P represents the probability that an event will occur, here the probability of owning a house, it is measured by the area of the standard normal curve from -∞ as shown in Figure 3.3.1.



Probit model: (a) given I_i, read P_i from the ordinate; (b) given P_i, read I_i from the abscissa.

Fig. 3.3.1. Probit model

Now to obtain information, in I_i the utility index, as well asβ_1

and β_2, we take the inverse of (2) to obtain.

$$I_i = F^{-1}(I_i) = F^{-1}(P_i)$$

$$= \beta_i + \beta_2 X_i \quad \text{.....(4)}$$

Where F^{-1} is the inverse of the normal CDF. What all this means can be made clear from Figure in panel a of this figure we obtain from the ordinate the (cumulative) probability of owning a house given $I^*_i \leq I_i$ whereas in panel b we obtain from the abscissa the value of I_i Given the value of P_i which is simply the reverse of the former.

In the logit model the dependent variable is the log of the odds ratio, which is a linear function of the regressors. The probability function that underlies the logit model is the logistic distribution. If the data are available in grouped form, we can use OLS to estimate the parameters of the logit model, provided we take into account explicitly the heteroscedastic nature of the error term. If the data are available at the individual, or micro, level, nonlinear-in-the-parameter estimating procedures are called for. If we choose the normal distribution as the appropriate probability distribution, then we can use the probit model. This model is mathematically a bit difficult as it involves integrals. But for all practical purposes, both logit and probit models give similar results. In practice, the choice therefore depends on the ease of computation, which is not a serious problem with sophisticated statistical packages that are now readily available.

Logit Vs Probit Model

In most applications the logit and probit models are quite similar, the main difference being that the logistic distribution has slightly flatter tails, which can be seen from the below figure. That is to say, the conditional probability Pi approaches 0 or 1 at a slower rate in logit than in probit. Therefore, there is no compelling reason to choose one over the other. In practice many researchers choose the logit model because of its comparative mathematical simplicity. However it is to be noted that, the results of these two models are not directly comparable. So, one must be very cautious while interpreting the results of these models.

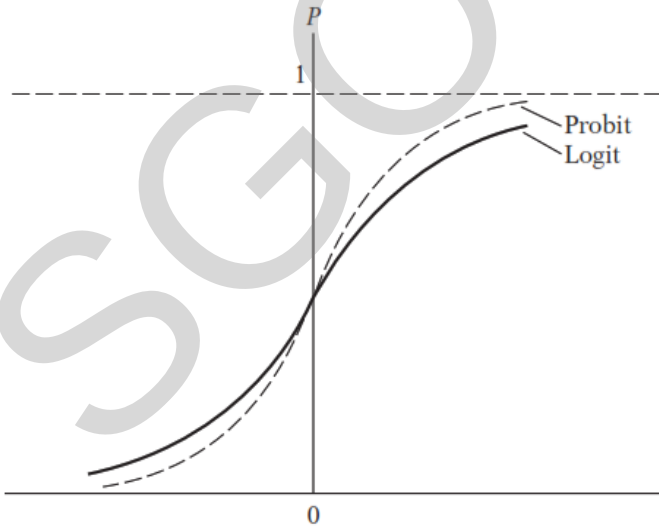


Fig. 3.3.2 Logit vs Probit Model

3.3.3.3 Tobit Model

In many data sets the dependent variable is continuous but is only observed above (or below) a certain limit.

• Examples

- Labor supply: hours Worked = 0 for the non-employed, positive otherwise.

- Consumer demand: expenditure on alcohol, durable-goods ownership, etc.

  If you run OLS on the positive observations only, you ignore the fact that zeros are informative ("non-participants").

  If you run OLS on the full sample treating zeros as true zeros, you under-estimate the slope because part of the variation is censored.

- We need a model that combines the intensive margin (how much) with the extensive margin (whether at all).

Basic idea

Postulate an unobserved latent variable yi^* that obeys the usual linear model

$yi^*=xi'\beta+\varepsilon i, \varepsilon i \sim N(0,\sigma_2)$.

What we actually see, yi, is "censored" at a threshold c (usually 0) :

$yi = \{yi^* \text{ if } yi^* > c, \text{otherwise.}$

- Left-censoring at 0 is the classic "Tobit I".
- Right-censoring or two-sided censoring are easy generalizations.

Interpreting Tobit results

1. Report three marginal effects: on the probability of being uncensored, on the conditional mean (given uncensored), and on the unconditional mean.
2. Translate coefficients into economically meaningful changes (e.g., "a $1 000 increase in income raises expected alcohol spending by $3 for the overall population, of which $1.50 comes from higher participation and $1.50 from higher intensity among drinkers").

Summarised Overview

Qualitative response models address settings where the dependent variable is categorical rather than continuous, posing distinctive challenges for estimation and interpretation. The Linear Probability Model (LPM) applies OLS to a binary outcome $Y \in \{0,1\}$ (for example, home ownership) with $E(Yi|Xi)=\beta_1+\beta_2 X_i$ interpreted as a conditional probability. While simple and transparent, the LPM has drawbacks: the error term is Bernoulli (hence non-normal), errors are heteroscedastic with $Var(ui)=P_i(1-P_i)$, and fitted probabilities can fall outside [0,1].

Weighted least squares using two-step estimated weights $\hat{w}_i = \hat{P}_i(1-\hat{P}_i)$ can address heteroscedasticity, and large samples mitigate non-normality, but the range and linearity issues persist. Logit and probit models remedy these by mapping a linear index to [0,1] via cumulative distribution functions: the logistic CDF for logit and the standard normal CDF for probit. In logit, the log-odds $\log\{P_i/(1-P_i\}$ are linear in X, ensuring probabilities

remain within bounds and permitting non-linear marginal effects that taper near 0 and 1; probit is analogous but uses the inverse normal link.

In practice, both typically yield similar substantive conclusions, with logistic tails slightly fatter; choice often hinges on convenience and interpretability. When the outcome is censored rather than strictly binary e.g., hours worked observed as zero for non-participants and positive otherwise the Tobit model introduces a latent continuous variable $y_i^* = x_i'\beta + \varepsilon_i$ observed only above (or below) a threshold. Naïvely running OLS on positives discards information, while treating zeros as true zeros biases slopes towards zero; Tobit jointly models participation (extensive margin) and intensity (conditional mean among the uncensored). Proper interpretation reports marginal effects on the probability of being uncensored, on the conditional mean given uncensored, and on the unconditional mean for the whole population, enabling economically meaningful statements about how covariates shift both participation and intensity.

Assignments

1. Distinguish qualitative-response models from continuous-outcome regressions and explain why OLS is inadequate in many binary/limited-dependent-variable settings.
2. Describe the data-generating processes that produce binary, ordered, and censored outcomes.
3. Use maximum likelihood (ML) or limited-information methods to estimate nonlinear models and assess convergence diagnostics.
4. Interpret estimated coefficients in terms of probabilities, marginal effects, elasticities, and latent variables.
5. Define a qualitative/limited dependent variable and give two real-world examples for which LPM might be an acceptable first pass.
6. Explain why fitted probabilities from an LPM can fall outside the [0,1] interval, yet OLS still yields unbiased slope estimates under strict exogeneity.
7. State the link functions used in Logit and Probit models and sketch their shapes on the same axes. Which one has fatter tails?

Reference

1. Damodar N Gujarati and Dawn C Porter (2009): *Basic Econometrics*, Fifth Edition, McGraw Hill International Edition.
2. Jeffrey M Wooldridge (2018): *Introductory Econometrics: A Modern Approach*, 7 th Edition, Thomson South Western.

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

BLOCK 4

Time Series and Panel Data Econometrics

UNIT 1

Time Series Properties and Autocorrelation

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the differences between stationary and non-stationary processes.
- understand the structure and implications of the random walk model
- learn how to compute and interpret autocorrelation and partial autocorrelation functions
- use these tools to prepare time series data for more advanced modeling (such as ARIMA)

Background

Economic and social phenomena, unlike many controlled laboratory processes, unfold over time and display patterns that change with evolving circumstances. To study such dynamic behaviour, economists and statisticians rely on time series analysis, which provides tools for understanding, modelling, and forecasting data observed periodically. Central to this analysis is the concept of stationarity - the idea that a time series should have statistical properties, such as mean and variance, that remain stable over time. Yet, in reality, many economic and financial series are non-stationary, often trending or exhibiting shocks that persist into the future.

One of the simplest yet most powerful illustrations of non-stationarity is the random walk model, a process in which current values depend heavily on past realizations plus a random disturbance.

To uncover underlying patterns and to determine how best to model a time series, researchers turn to measures of dependence such as the autocorrelation function (ACF) and the partial autocorrelation function (PACF). These tools help identify whether a series is driven by short-term dependencies, long-run trends, or essentially unpredictable shocks.

This unit introduces these foundational concepts - stationarity and non-stationarity, random walks, and the role of correlation structures - laying the groundwork for advanced techniques in time series and forecasting. Understanding these ideas is crucial for anyone seeking to interpret dynamic data in economics, finance, or the social sciences.

Keywords

Time Series Data, Stationarity, Non-Stationarity, Unit Root, Random Walk, Autocorrelation (ACF), Partial Autocorrelation (PACF)

Discussion

4.1.1 Time Series Analysis

A time series is a sequence of data points, measured typically at successive points in time spaced at uniform time intervals. Time series data have a natural temporal ordering. This makes time series analysis distinct from cross-sectional studies, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their respective education levels, where the individuals' data could be entered in any order). Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely related than observations further apart.

In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

Time series analysis can be applied to real-valued, continuous data, discrete numeric data, or discrete symbolic data (i.e. sequences of characters, such as letters and words in the English language.)

Stochastic Process: A random or stochastic process is a collection of random variables ordered in time.

Stationary Stochastic Process: (SSP) A type of stochastic process that has received a great deal of attention and scrutiny by time series analyses is SSP.

A stochastic process is said to be state if its means and variance are constant over time and the value of the Cov. between the two time period depends only on the distance between the two time period.

4.1.2 Stationarity, Non-Stationarity

Strict Stationarity: A time series $\{X_t\}$ is strictly stationary if the joint distribution of $(X_{\{t_1\}}, X_{\{t_2\}}, \ldots, X_{\{t_k\}})$ is identical to the joint distribution of $(X_{\{t_1+h\}}, X_{\{t_2+h\}}, \ldots, X_{\{t_k+h\}})$ for all $t_1, t_2, \ldots, t_k$ and h. A time series is strictly stationary if its statistical properties do not change over time.

Weak Stationarity (Covariance Stationarity): A time series $\{X_t\}$ is weakly stationary if:

1. $\boldsymbol{E(X_t) = \mu}$ (constant mean)

2. $\boldsymbol{Var(X_t) = \sigma^2}$ (constant variance)

3. $\boldsymbol{Cov(X_t, X_{\{t-k\}}) = \gamma(k)}$ (autocovariance depends only on lag k)

A test of stationarity that has become widely popular over the past several years is unit root test.

$$Y_t = \rho Y_{t-1} + U_t \quad -1 \leq \rho \leq 1 \ldots\ldots (1)$$

When ρ =1 (1) becomes a random walk model unit drift.

Subtract Y_{t-1} from both the sides of the equation

$$Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + u_t$$

$$= (\rho - 1)\, Y_{t-1} + u_t \text{----------(2)}$$

$$\Delta Y_t = \delta Y_{t-1} + u_2 \text{----------(3)}$$

(Using 1st differential operate)

δ =0, if δ = 0 then ρ = 1

that is we have unit root, time series under consideration is nonstationary.

$$\Delta Y_t = (Y_t - Y_{t-1}) = u_t \text{----------(4)}$$

The term non-stationary, random walk and unit root can be treated as synonymous.

Non-Stationarity

If a time series is not stationary it is called non-stationary time series. RWM is a classical example. It is said that asset purchased such as stock purchases follows a random walk i.e. they are non- stationary.

Random Walk (Unit Root Process Basics)

A random walk means:

$$Y_t = \delta Y_{t-1} + u_t$$

Here u_t = random shock (white noise).

If δ=1, then Yt depends fully on the previous value plus shock.

This makes the process non-stationary (variance grows over time).

There are two types:

Random walk without drift (i.e. no constant)

Random walk with drift (i.e. a constant term)

4.1.3 Random Walk Models

In discussing the nature of the unit root process, we noted that a random walk process may have no drift, or it may have drift or it may have both deterministic and stochastic trends. To allow for the various possibilities, the DF test is estimated in three different forms, that is, under three different null hypotheses.

***Yt* is a random walk:** $Y_t = \delta Y_{t-1} + u_t$

***Yt* is a random walk with drift:** $Y_t = \beta_1 + \delta Y_{t-1} + u_t$ **around a stochastic trend**:

$$Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + u_t$$

where t is the time or trend variable. In each case, the null hypothesis is that δ = 0; that is, there is a unit root - the time series is nonstationary. The alternative hypothesis is that δ is less than zero; that is, the time series is stationary. If the null hypothesis is rejected, it means that Y_t is a stationary time series with zero mean in the case of (2), that Y_t is stationary with a nonzero mean [= â1/(1 − ñ)] in the case of (4) . It is extremely important to note that the critical values of the tau test to test the hypothesis that ä = 0, are different for each of the preceding three specifications of the DF test. Moreover, if, say, specification (4) is correct, but we estimate (2), we will be committing a specification error, The same is true if we estimate (4) rather than the true (5). Of course, there is no way of knowing which specification is correct to begin with. Some trial and error are inevitable, data mining notwithstanding. The actual estimation procedure is as follows: Estimate (2), or (4) by OLS; divide the estimated coefficient of Y_{t-1} in each case by its standard error to compute the (ô) tau statistic; and refer to the DF tables (or any statistical package). If the computed absolute value of the tau statistic (|ô |) exceeds the DF or MacKinnon critical tau values, we reject the hypothesis that δ = 0, in which case the time series is stationary. On the other hand, if the computed |τ | does not exceed the critical tau value, we do not reject the null hypothesis, in which case the time series is nonstationary. Make sure that you use the appropriate critical τ values.

Random Walk Without Drift

$Y_t - Y_{t-1} + u_t$----------(1)

U_t = white nose.

Y_t = RW

We can write eq. (1) as $Y_1 = Y_0 + u_1$

$Y_2 = Y_1 + u_2 = Y_0 + u_1 + u_2$

$Y_3 = Y_2 + u_3 = Y_0 + u_1 + u_2 + u_3$

$Y_t = Y_0 + \Sigma u_1$----------(2)

$\therefore E(Y_t) = E(Y_0 + \Sigma u_t) = Y_0$----------(6)

Var $Y_t = t\sigma^2$

RWM without drift is a non stationary stochastic process.

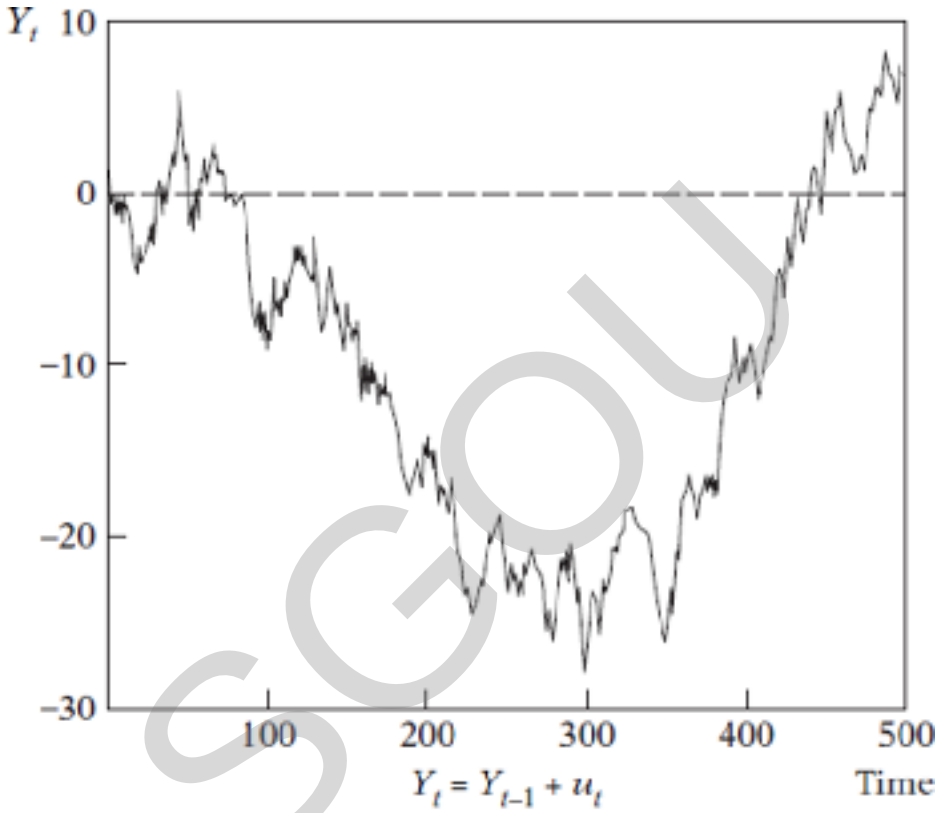


Fig. 4.1.1 A Random Walk without Drift

$$Y_t = Y_{t-1} + u_t$$

Random Walk With Drift

$$Y_t = \delta + Y_{t-1} + u_t(1)$$

δ = drift parameter

$$Y_t + Y_{t-1} = DY_t = \delta + u_t$$----------(2)

Yt = upward or downward depending on δ being +ve or -ve

$E(Y_t) = Y_0 + t\delta$---------(3)

Var $(Y_t) = = t\ \sigma^2$----------(4)

A Random Walk Model (RWM) with drift has a variance that increases over time, thereby violating the condition of (weak) stationarity. In other words, an RWM with drift is a non-stationary stochastic process. The RWM is an example of a unit root process

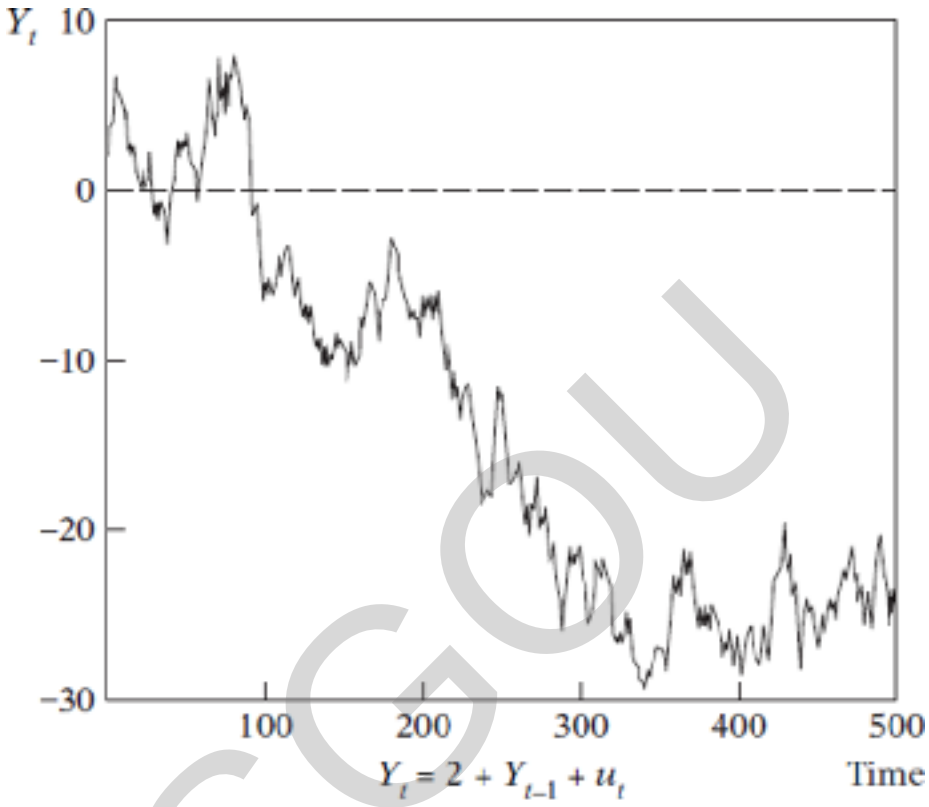


Fig. 4.1.2 A Random Walk with Drift

Simple Random Walk: $X_t = X_{\{t-1\}} + \varepsilon_t$

This is the prototypical unit root process where ρ = 1.

Random Walk with Drift: $X_t = \alpha + X_{\{t-1\}} + \varepsilon_t$

Where α represents the drift parameter.

Random Walk with Drift and Trend: $X_t = \alpha + \beta t + X_{\{t-1\}} + \varepsilon_t$

Where β is the deterministic trend coefficient.

4.1.4 Autocorrelations

Some of the possible reasons for the introduction of autocorrelation in the data are as follows:

1. Carryover of effect, at least in part, is an important source of autocorrelation. For example, the monthly data on expenditure on household is influenced by the expenditure of preceding month. The autocorrelation is present in cross-section data as well as time-series data. In the cross-section data, the neighbouring units tend to be similar with respect to the characteristic under study. In time-series data, time is the factor that produces autocorrelation. Whenever some ordering of sampling units is present, the autocorrelation may arise.

2. Another source of autocorrelation is the effect of deletion of some variables. In regression modelling, it is not possible to include all the variables in the model. There can be various reasons for this, e.g., some variable may be qualitative, sometimes direct observations may not be available on the variable etc. The joint effect of such deleted variables gives rise to autocorrelation in the data.

3. The misspecification of the form of relationship can also introduce autocorrelation in the data. It is assumed that the form of relationship between study and explanatory variables is linear. If there are log or exponential terms present in the model so that the linearity of the model is questionable, then this also gives rise to autocorrelation in the data.

4. The difference between the observed and true values of the variable is called measurement error or errors–in-variable. The presence of measurement errors on the dependent variable may also introduce the autocorrelation in the data.

The following structures are popular in autocorrelation:

1. Autoregressive (*AR*) process.
2. Moving average (*MA*) process.
3. Joint autoregression moving average (*ARMA*) process.

Estimation under the First Order Auto Regressive Process

A first-order autoregressive process, or AR(1) process, is a time series model where the current value of a variable is a linear function of its immediately preceding value, plus a constant and a random error term.

Let Y_t represent the lagged GDP at time *t*. If we model Y_t as

$$Y_t - \delta = \alpha_1(Y_{t-1} - \delta) + u_t$$

where δ is the mean of Y and where ut is an uncorrelated random error term with zero mean and constant variance σ 2 (i.e., it is white noise), then we say that Yt follows a first-order autoregressive, or AR(1), stochastic process. Here the value of Y at time t depends on its value in the previous time period and a random term; the Y values are expressed as deviations from their mean value. In other words, this model says that the forecast value of Y at time t is simply some proportion (- α1) of its value at time (t - 1) plus a random shock or disturbance at time t; again the Y values are expressed around their mean values.

But if we consider this model,

$Y_t - \delta = \alpha_1(Y_{t-1} - \delta) + \alpha_3(Y_{t-2} - \delta) + u_t$

then we say that Yt follows a second-order autoregressive, or AR (2), process. That is, the value of Y at time t depends on its value in the previous two time periods, the Y values being expressed around their mean value δ.

In general, we can have

$Y_t - \delta = \alpha_1(Y_{t-1} - \delta) + \alpha_2(Y_{t-2} - \delta) + \text{----------} + \alpha_p(Y_{t-p} - \delta) + u_t$

in which case Yt is a pth-order autoregressive, or AR(p), process.

Notice that in all the preceding models only the current and previous Y values are involved; there are no other regressors. In this sense, we say that the "data speak for them- selves." They are a kind of reduced form model.

Tests for Autocorrelation

Durbin Watson test:

Autocorrelation, also known as serial correlation is the correlation of a signal with a delayed copy of itself as a function of delay. In other words, it is the similarity between observations as a function of the time lag between them. This property makes the DW test useful for time-series data where the current state of the system depends heavily on prior data. Time-series refers to systems in which time is a factor in the progression of the system, for example: ocean tides, sunspot counts, stock-market prices, etc.

The process is basically a linear regression of the data in the current series against one or more past values in the same series. The value of the outcome variable (Y) at some point t in time is - like "regular" linear regression - directly related to the predictor variable (X). Where simple linear regression and autoregression models differ is that Y is dependent on X and previous values for Y. This an example of a stochastic process, which has degrees of uncertainty or randomness built in. The randomness means that you might be able to predict future trends pretty well with past data, but you're never going to get 100 percent accuracy. Usually, the process gets "close enough" for it to be useful in most scenarios.

The DW test is named after James Durbin and Geoffrey Watson who first used the technique in the 50's. In statistics, the autocorrelation of a random process is the Pearson correlation between values of the process at different times, as a function of the two times or of the time lag. The Durbin-Watson Test is a measure of autocorrelation (also called serial correlation) in residuals from regression analysis. The DW test statistic is calculated using the following equation:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

where T is the total number of observations and e is the residual error given by:

$$e_t = \rho e_{t-1} + \nu_t$$

where rho is the sample autocorrelation of the residuals. From the second equation, one can see how the current residuals are related to the past state of the system. If one has a lengthy sample, then this can be linearly mapped to the Pearson correlation of the time-series data with its lags. It must be noted that the DW test is used to test for a specific type of autoregression (AR1), meaning a first order autoregression. The outcome variable in a first order autoregression process at some point in time t is only related to time periods that are one period apart (i.e. the value of the variable at t-1). A second or third order autoregression process would be related to data two or three periods apart respectively.

The hypotheses for the Durbin Watson test are:

H0 = no first order autocorrelation (rho = 0)

H1 = first order autocorrelation exists (rho != 0)

One assumes:

That the errors are normally distributed with a mean of 0.

The errors are stationary.

The Durbin-Watson test gives values that are between 0 and 4 with the following meaning:

2 is no autocorrelation.

0 to <2 is positive autocorrelation (common in time series data).

2 to 4 is negative autocorrelation (less common in time series data).

Once again, the test statistic is approximately equal to 2*(1-rho) where rho is the sample autocorrelation of the residuals. Thus, for rho = 0, indicating no serial correlation, the test statistic equals 2. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation.

Upper and lower critical values, dU and dL have been tabulated for different values of k (the number of explanatory variables) and n.

If d<dL — reject H0

If d>dU — do not reject H0

If dL < d < dU — test is inconclusive

Using the following DW critical value table, by inputting sample size n, number of regressors k and acceptable alpha level one can find dU and dL.

Table 4.1.1 Critical Values of Durbin - Watson Statistic

BOUNDS FOR CRITICAL VALUES OF THE DURBIN-WATSON STATISTIC

1% SIGNIFICANCE POINTS OF Q_L AND Q_U

	$\Lambda = 2$		$\Lambda = 3$		$\Lambda = 4$		$\Lambda = 5$		$\Lambda = 6$	
N	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U
15	0.811	1.069	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967
16	0.844	1.087	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587
55	1.356	1.428	1.320	1.466	1.284	1.505	1.246	1.548	1.209	1.592
60	1.382	1.449	1.351	1.484	1.317	1.520	1.283	1.559	1.248	1.598
65	1.407	1.467	1.377	1.500	1.346	1.534	1.314	1.568	1.283	1.604
70	1.429	1.485	1.400	1.514	1.372	1.546	1.343	1.577	1.313	1.611
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.340	1.617
80	1.465	1.514	1.440	1.541	1.416	1.568	1.390	1.595	1.364	1.624
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.630
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.641
100	1.522	1.562	1.502	1.582	1.482	1.604	1.461	1.625	1.441	1.647

BOUNDS FOR CRITICAL VALUES OF THE DURBIN-WATSON STATISTIC

2.5% SIGNIFICANCE POINTS OF Q_L AND Q_U

	$\Lambda = 2$		$\Lambda = 3$		$\Lambda = 4$		$\Lambda = 5$		$\Lambda = 6$	
N	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U
15	0.949	1.222	0.827	1.405	0.706	1.615	0.588	1.848	0.478	2.099
16	0.980	1.235	0.864	1.403	0.748	1.594	0.636	1.806	0.527	2.035
17	1.009	1.248	0.899	1.403	0.788	1.578	0.680	1.773	0.574	1.983
18	1.035	1.261	0.930	1.405	0.825	1.567	0.720	1.746	0.619	1.939
19	1.060	1.274	0.959	1.407	0.859	1.558	0.758	1.724	0.660	1.902
20	1.082	1.286	0.988	1.410	0.890	1.551	0.794	1.705	0.699	1.871
21	1.104	1.297	1.012	1.415	0.920	1.546	0.826	1.691	0.734	1.845
22	1.124	1.308	1.036	1.419	0.947	1.543	0.858	1.678	0.769	1.823
23	1.144	1.319	1.059	1.424	0.973	1.541	0.887	1.668	0.801	1.804
24	1.161	1.329	1.080	1.429	0.997	1.539	0.914	1.659	0.830	1.787
25	1.178	1.339	1.099	1.435	1.019	1.539	0.939	1.652	0.859	1.773
26	1.194	1.348	1.118	1.439	1.041	1.538	0.964	1.646	0.886	1.761
27	1.208	1.358	1.135	1.445	1.061	1.539	0.986	1.641	0.911	1.751
28	1.222	1.367	1.153	1.450	1.080	1.540	1.007	1.637	0.934	1.742
29	1.236	1.375	1.168	1.455	1.098	1.541	1.028	1.634	0.958	1.734
30	1.249	1.383	1.183	1.460	1.115	1.542	1.047	1.632	0.978	1.727
31	1.261	1.391	1.197	1.465	1.132	1.544	1.066	1.630	0.999	1.721
32	1.273	1.399	1.211	1.469	1.147	1.546	1.083	1.628	1.018	1.715
33	1.284	1.406	1.224	1.474	1.163	1.548	1.099	1.627	1.037	1.711
34	1.294	1.413	1.236	1.479	1.176	1.550	1.115	1.626	1.054	1.707
35	1.305	1.420	1.248	1.484	1.190	1.553	1.131	1.626	1.071	1.704
36	1.315	1.426	1.259	1.488	1.203	1.555	1.145	1.625	1.087	1.701
37	1.324	1.433	1.270	1.493	1.215	1.557	1.159	1.625	1.102	1.698
38	1.333	1.439	1.281	1.497	1.227	1.560	1.173	1.625	1.117	1.695
39	1.342	1.445	1.291	1.501	1.238	1.562	1.185	1.626	1.131	1.693
40	1.350	1.451	1.300	1.506	1.249	1.564	1.197	1.626	1.144	1.692
45	1.388	1.477	1.343	1.525	1.298	1.576	1.252	1.630	1.204	1.687
50	1.420	1.500	1.380	1.543	1.338	1.588	1.297	1.636	1.255	1.685
55	1.447	1.520	1.411	1.559	1.373	1.600	1.335	1.642	1.297	1.686
60	1.471	1.538	1.438	1.573	1.404	1.610	1.369	1.649	1.333	1.688
65	1.492	1.554	1.461	1.587	1.430	1.620	1.398	1.655	1.365	1.691
70	1.511	1.568	1.482	1.599	1.453	1.630	1.424	1.662	1.393	1.695
75	1.528	1.582	1.501	1.610	1.474	1.638	1.446	1.668	1.418	1.699
80	1.543	1.594	1.518	1.619	1.493	1.647	1.467	1.674	1.441	1.703
85	1.557	1.605	1.534	1.629	1.510	1.654	1.485	1.680	1.461	1.707
90	1.570	1.614	1.548	1.638	1.525	1.662	1.502	1.686	1.479	1.711
95	1.582	1.624	1.560	1.646	1.539	1.668	1.517	1.691	1.495	1.715
100	1.593	1.633	1.573	1.654	1.552	1.675	1.532	1.696	1.511	1.718

BOUNDS FOR CRITICAL VALUES OF THE DURBIN-WATSON STATISTIC

5% SIGNIFICANCE POINTS OF Q_L AND Q_U

	$\Lambda = 2$		$\Lambda = 3$		$\Lambda = 4$		$\Lambda = 5$		$\Lambda = 6$	
N	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U	Q_L	Q_U
15	1.077	1.361	0.945	1.543	0.814	1.750	0.685	1.977	0.562	2.220
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104
18	1.158	1.392	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060
19	1.180	1.401	1.075	1.535	0.967	1.685	0.859	1.848	0.752	2.022
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.828	1.964
22	1.240	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.786	0.895	1.919
24	1.273	1.446	1.188	1.546	1.101	1.657	1.013	1.775	0.925	1.902
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873
27	1.316	1.468	1.240	1.556	1.162	1.651	1.083	1.753	1.004	1.861
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.070	1.833
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825
32	1.373	1.502	1.309	1.573	1.244	1.650	1.177	1.732	1.109	1.819
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813
34	1.393	1.514	1.332	1.580	1.271	1.652	1.208	1.728	1.144	1.807
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803
36	1.411	1.524	1.354	1.587	1.295	1.654	1.236	1.725	1.175	1.799
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795
38	1.427	1.535	1.373	1.594	1.317	1.656	1.261	1.723	1.204	1.792
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.231	1.786
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.334	1.771
55	1.527	1.601	1.490	1.640	1.452	1.681	1.414	1.724	1.374	1.768
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767
70	1.583	1.641	1.554	1.671	1.524	1.703	1.494	1.735	1.464	1.768
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.486	1.770
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772
85	1.624	1.671	1.600	1.696	1.575	1.721	1.551	1.747	1.525	1.774
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780

The test procedure is as follows:

$H_0 : \rho = 0$			
Nature of H_1	Reject H_0 when	Retain H_0 when	The test is inconclusive when
$H_1 : \rho > 0$	$d < d_L$	$d > d_U$	$d_L < d < d_U$
$H_1 : \rho < 0$	$d > (4 - d_L)$	$d < (4 - d_U)$	$(4 - d_U) < d < (4 - d_L)$
$H_1 : \rho \neq 0$	$d < d_L$ or $d > (4 - d_L)$	$d_U < d < (4 - d_U)$	$d_L < d < d_U$ or $(4 - d_U) < d < (4 - d_L)$
Values of d_L and d_U are obtained from tables.			

Limitations of D-W Test

1. If d falls in the inconclusive zone, then no conclusive inference can be drawn. This zone becomes fairly larger for low degrees of freedom. One solution is to reject H_0 if the test is inconclusive. A better solutions is to modify the test as

   - Reject H_0 when $d < d_U$.
   - Accept H_0 when $d \geq d_U$.

This test gives a satisfactory solution when values of x_i 's change slowly, e.g., price, expenditure etc.

2. The *D-W* test is not applicable when the intercept term is absent in the model. In such a case, one can use another critical value, say d_M in place of d_L . The tables for critical values d_M are available.

3. The test is not valid when lagged dependent variables appear as explanatory variables. For example,

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_r y_{t-r} + \beta_{r+1} x_{t1} + ... + \beta_k x_{t,k-r} + u_t ,$$

$$u_t = \rho u_{t-1} + \varepsilon_t .$$

Estimation procedures when autocorrelation co-efficient is known:

Consider the estimation of regression coefficient under first-order autoregressive disturbances and the autocorrelation coefficient is known. The model is

$y = X\beta + u,$

$u_t = \rho u_t + \varepsilon_t$

and assume that $E(u) = 0,\ E(uu') = \psi \neq \sigma^{2)} I,\ E(\varepsilon) = 0,\ E(\varepsilon\varepsilon') = \sigma^2 I.$

The OLSE of β is unbiased but not, in general, efficient, and the estimate of σ^2 generalized least squares estimation procedure, and GLSE of β is

$\hat{\beta} = (X'\psi^{-1}X)^{-1}X'\psi^{-1}y$ where is biased. So we use

$$\begin{bmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & -\rho & \dots & 1 \end{bmatrix}$$

To employ this, we proceed as follows:

1. Find a matrix P such that $P'P = \psi^{-1}$. In this case

$$\begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \dots & 0 \\ -\rho & 1 & 0 & \dots & 0 \\ 0 & -\rho & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

2. Transform the variables as

$y^* = Py,\ X^* = PX,\ \varepsilon^* = P\varepsilon$.

If the first column of X is a vector of ones, then the first column of X^* is not constant. Its first element is $\sqrt{1-\rho^2}$.

Now employ OLSE with observations

$\beta^* = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$, its covariance matrix is

$V(\hat{\beta}) = \sigma^2(X^{*\prime}X^*)^{-1}$

$= \sigma^2(X'\psi^{-1}X)^{-1}$ and its estimator is

$\hat{V}(\hat{\beta}) = \hat{\sigma}^2(X'\psi^{-1}X)^{-1}$

y^* and X^*, then the OLSE of β is

where

$$= \frac{\hat{\sigma}^2\ (y - X\hat{\beta})'\psi^{-1}(y - X\hat{\beta})}{n-k}$$

Estimation procedures when autocorrelation co-efficient is unknown:

Several procedures have been suggested to estimate the regression coefficients when autocorrelation coefficient is unknown. The feasible GLSE of β is

$\hat{\beta} = (X' \hat{\Omega}^{-1} X)^{-1} X' \hat{\Omega}^{-1} y$

where $\hat{\Omega}^{-1}$ is the Ψ^{-1} matrix with ρ replaced by its estimator $\hat{\rho}$.

1. Use of simple correlation co-efficient

The most common method is to use the sample correlation coefficient r between successive residuals as the natural estimator of ρ.

2. Durbin procedure

In Durbin procedure, the model

$y_t - \rho y_{t-1} = \beta_0 (1-\rho) + \beta (x_t - \rho x_{t-1}) + \varepsilon_t$, $t = 2,3,...,n$

is expressed as

$y_t = \beta_0 (1-\rho) + \rho y_{t-1} + \beta x_1 - \rho\beta x_{t-1} + \varepsilon_t$

$= \beta^* + \rho y \quad + \beta x + \beta^* x \quad + \varepsilon$, $t = 2,3,...,n$ (*)

0 $t-1$ t $t-1$ t

where $\beta^* = \beta (1-\rho)$, $\beta^* = -\rho\beta$.

Now run a regression using OLS to model (*) and estimate r^* as the estimated coefficient of y_{t-1}.

Another possibility is that since $\rho \in (-1,1)$, so search for a suitable ρ which has smaller error sum of squares.

3. Cochrane-Orcutt Procedure

This procedure utilizes P matrix defined while estimating β when ρ is known. It has following steps:

i. Apply OLS to $y_t = \beta_0 + \beta_1 x_t + u_t$ and obtain the residual vector e.

$\sum e_t e_{t-1}$

ii. Estimate ρ by $r =$

Note that r is a consistent estimator of ρ.

iii. Replace ρ by r is

$y_t - \rho y_{t-1} = \beta_0 (1-\rho) + \beta (x_t - \rho x_{t-1}) + \varepsilon_t$

and apply OLS to the transformed model

$y - ry = \beta^* + \beta (x - rx) +$ disturbance term

t $t-1$ 0 t $t-1$

and obtain estimators of β_0^* and β as $\hat{\beta}_0^*$ and $\hat{\beta}$ respectively.

This is Cochrane-Orcutt procedure. Since two successive applications of OLS are involved, so it is also called as **two-step procedure.**

This procedure is repeated until convergence is achieved, i.e., iterate the process till the two successive estimates are nearly same so that stability of estimator is achieved.

This is an iterative procedure and is numerically convergent procedure. Such estimates are asymptotically efficient and there is a loss of one observation.

4.1.6 Partial Autocorrelation Functions

Autocorrelation is not the correct measure of the mutual correlation between Yt and Yt+k in the presence of the intermediate variables. Partial autocorrelation solves this problem by measuring the correlation between Yt and Yt+k when the influence of the intermediate variables has been removed. Hence partial autocorrelation in time series analysis defines the correlation between Yt and Yt+k which is not accounted for by lags t +1 to t + k –1. The partial autocorrelation function is similar to the atocorrelation function except that it displays only the correlation between two observations after removing the effect of intermediate variables. For example, if we are interested in the direct relationship between today's consumption of patrol and that of a year ago then we don't blame what happens in between. The consumption of the previous 12 months has an effect on the consumption of the previous 11 months, and the cycle continues until the most current period. In partial autocorrelation estimates, these indirect effects are ignored. Therefore, we can define the partial autocorrelation function as The partial autocorrelation function calculates the degree of relationship between a time series Yt with its own lagged values Yt+k after their mutual linear dependency on the intervening variables Yt+1, Yt+2,…, Yt+k–1 has been removed.

We can define the partial autocorrelation function between Yt and Yt+k as the conditional correlation between Yt and Yt+k, conditional on Y ,Y ,...,Y t1 t2 tk1 + + +– (the set of observations that come between the time points Yt and Yt+k), is known as the kth order PACF.

This is the correlation between values that are two time periods apart conditional on knowledge of the value in between. (By the way, the two variances in the denominator will equal each other in a stationary series.), therefore,

$$\phi_{kk} = \frac{\mathrm{Cov}\left(Y_t, Y_{t+k} \mid Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}\right)}{\mathrm{Var}\left(Y_t \mid Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}\right)}$$

The formula for calculating the partial autocorrelation function looks scary, therefore, we calculate it using the autocorrelation function instead of it. The 1st order partial autocorrelation function equals to the 1st order autocorrelation function, that is,

$$\phi_{11} = \rho_1$$

Similarly, we can define the 2nd order (lag) partial autocorrelation function in terms of autocorrelation function as

$$\phi_{22} = \frac{(\rho_2 - \rho_1^2)}{(1 - \rho_1^2)}$$

Summarised Overview

Time series analysis begins with the vital task of diagnosing whether a dataset is stationary - meaning its statistical properties such as mean and variance remain constant over time. Stationarity is crucial because many econometric models and forecasting techniques rely on this assumption for credible results; non-stationary data can lead to misleading conclusions and unreliable predictions. Methods like differencing and unit root tests help transform or identify non-stationary series for proper modelling.

Random walk models exemplify non-stationarity, capturing processes in which shocks persist and where future values depend heavily on previous ones plus unpredictable, random influences. Recognizing when a process follows a random walk is essential for choosing the correct forecasting approach.

Autocorrelation and partial autocorrelation functions (ACF and PACF) empower analysts to diagnose the temporal dependence structure of time series data. The ACF measures the overall correlation at various lags, while the PACF narrows in on the direct effect of each specific lag after removing the influence of earlier lags - guiding model specification, particularly for autoregressive (AR) and ARIMA models.

Assignments

1. Distinguish between mean stationarity and covariance stationarity with examples
2. What is a random walk process? Illustrate with a simple equation and discuss its properties.
3. Explain the consequences of fitting an OLS regression to non-stationary time series data.
4. Given the following sample autocorrelations: ρ1=0.7, ρ2 =0.5, ρ3=0.1, estimate the first two values of the partial autocorrelation function (PACF

Reference

1. Walter Enders (2013) *Applied Econometric Time Series*, 3rd Edition, Wiley, ISBN-10: 8126543914, ISBN-13: 978-8126543915.
2. Jeffrey M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition, MIT Press, ISBN-10: 9384106259, ISBN-13: 978-9384106256
3. James D Hamilton *Time Series Analysis* (1994) Princeton University Press

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 2

Unit Root Testing and ARIMA Models

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the concept of unit roots in time series data and their implications for stationarity and long-term predictability
- learn how to conduct and interpret unit root tests (such as the Dickey-Fuller and Augmented Dickey-Fuller [ADF] tests) to determine the presence of unit roots in economic and financial time series
- comprehend the characteristics, mathematical formulation, and model-building approaches for Autoregressive (AR) and Moving Average (MA) processes
- gain proficiency in constructing and interpreting ARMA (Autoregressive Moving Average) models that combine AR and MA components for modelling stationary time series

Background

Economic and financial time series often exhibit complex patterns driven by underlying trends, persistence, and random shocks. Before constructing reliable models and forecasts, it is essential to determine whether these series are stationary or possess unit roots—a feature that indicates non-stationarity and persistent effects from shocks. Identifying unit roots and testing for their existence (using methods like the Dickey-Fuller test) is a foundational step in modern time series econometrics, as it influences both theoretical interpretations and practical modelling strategies.

Once the presence or absence of unit roots is established, the analyst can select from a suite of modelling frameworks:

Autoregressive (AR) and Moving Average (MA) models capture direct dependence on past values and error terms, respectively.

Integrated processes (I(d)) describe series that achieve stationarity only after differencing d times, allowing analysts to transform non-stationary series for meaningful modelling.

ARMA and ARIMA models provide flexible structures that combine autoregressive, moving average, and integration components—widely used for forecasting, policy evaluation, and structural analysis in economics.

By mastering the concepts and diagnostic tools in this unit, researchers and policy analysts are equipped to discern whether shocks will fade or persist, choose parsimonious models, and generate robust forecasts rooted in the true nature of the data-generating process.

Keywords

Unit Root, Autoregressive (AR) Model, Moving Average (MA) Model, Integrated Process (I(d)), ARMA Model, ARIMA Model

Discussion

4.2.1 Unit Roots and Units Roots Tests

The concept of unit roots represents one of the most revolutionary developments in modern econometrics and time series analysis. The recognition that many economic time series are non-stationary - containing unit roots - fundamentally changed how economists approach empirical analysis and transformed our understanding of economic relationships. This paradigm shift began in the 1970s and 1980s with seminal work by economists like Dickey, Fuller, Nelson, and Plosser, who demonstrated that many macroeconomic variables follow random walks rather than stationary processes.

Understanding unit roots is crucial for several reasons. First, the presence of unit roots affects the statistical properties of estimators and test statistics in ways that can lead to seriously misleading inferences. Second, unit root tests are essential for determining the appropriate modelling strategy for time series data. Third, the distinction between stationary and non-stationary series is fundamental to concepts like cointegration and error correction models, which form the backbone of modern macroeconomic modelling.

This comprehensive guide will take you through the theoretical foundations, mathematical derivations, testing procedures, and practical applications of unit root analysis, providing you with the tools necessary to conduct rigorous empirical research with time series data.

Theoretical Foundations of Unit Roots

Unit Root Process

Definition: A series contains a unit root if it can be written as: $X_t = \rho X_{\{t-1\}} + \varepsilon_t$

where ρ = 1 and ε_t is a stationary error term.

Autoregressive Process of Order 1: $X_t = \rho X_{\{t-1\}} + \varepsilon_t$

Where ε_t ~ iid(0, σ²_ε)

Three Cases:

1. **|ñ| < 1**: Stationary process
2. **ñ = 1**: Unit root (non-stationary)
3. **|ñ| > 1**: Explosive process (non-stationary)

Properties of Unit Root Processes

Variance: For a simple random walk starting at X_0: $Var(X_t) = t\sigma_\varepsilon^2$

The variance increases linearly with time, violating the constant variance requirement for stationarity.

Autocovariance: $Cov(X_t, X_{\{t-k\}}) = (t-k)\sigma_\varepsilon^2$ for k < t

The autocovariance depends on both the lag k and the time t.

Persistence: Shocks to unit root processes have permanent effects: $\frac{\partial X_{\{t+j\}}}{\partial \varepsilon_t} = 1 \; for \; all \; j \geq 0$

This contrasts with stationary processes where shocks eventually die out.

Unit root is troublesome.

- For one thing, the law of large number (LLN) does not hold for a unit root process.
- For a stationary and ergodic process LLN states that as T → ∞1T∑Tt=1yt → E(yt)

Unit root may cause three troubles. First, E(yt) may not be a constant. Second, the variance of yt is non-constant. Third, the serial correlation between yt and yt−j decays to zero very slowly

Unit Root Test

A test of stationary that has become widely popular over the past several years as unit root test.

$Y_t = \rho Y_{t-1} + u_t \; -1 \leq \rho \leq 1 \quad (1)$

ρ =1

1) Become a random walk model unit drift. Sub Y_{t-1} from b/s

$Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + u_t$

$= (p-1)Y_{t-1} + u_t$ ----------(2)

$\Delta Y_t = \delta Y_{t-1} + u_2$ ----------(3)

(Using 1st differential operate)

$\delta = 0$, *if* $\delta = 0$ *then* $\rho = 1$, that is we have unit root, time series under consideration is non-stationary.

$\Delta Y_t = (Y_t - Y_{t-1}) = u_t$ ----------(4)

The term non-stationary, random walk and unit root can be treated as synonymous.

$|\rho| \leq |1$ i.e. the time serves It is stationary.

The time series variables included in regression models need to be stationary because if their means and variances are changing, the computed t-statistics under the OLS regression fail to converge to their true value as sample size increases. Although the variables have strong association between them although in reality there might not be any such association between the variables. This is known as the problem of Spurious regression.

Now start with the stationary stochastic process, a stochastic process (time series) Yt is said to be stationary if its mean and variance are constant and independent of time and the covariances depend only upon the disturbance between two time periods, but not on time periods per se. So, Yt is stationary when the following conditions hold:

i. E(Yt) = μ = constant for all t

ii. Var(Yt) = σ2 = Constant for all t

iii. Cov(Yt, Yt-s) = λs = Constant for all t ≠ s

These conditions imply that the mean and variance of the stationary series remain constant over time. For example, if we consider monthly observations from 2010 to 2020 then the above conditions remain same during this period.

How to check Stationarity problem in your data set?

There are many ways:

1. Graphical Approach: In general, stationarity of a series can be understood simply by plotting the series over time. If the series shows no tendency to drift upwards over time, it is stationary in mean. Otherwise, it is non-stationary.

2. Autocorrelation function (ACF) and Correlogram: The stationarity of a given time series may be assessed by computing the value of its autocorrelation function (ρ). For the series Yt, the value of ρ at lag k, denoted by ρk, is computed as under:

$$\rho_k = \frac{\lambda_k}{\lambda_0} = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)}$$

Here, we may compute the value of ρk for different lag lengths (k). A graphical plot of ρk against k is called correlogram. The correlogram helps to understand whether the series is stationary or not. If the value of ρk at various hovers/stay close around 0, we say that the series is stationary. For a non-stationary series, the value of ρk at various lags are non-zero, although they may slowly decline towards 0 as the lag length increases.

3. Bartlett Test: Here, it is shown that if the series is purely random, ρk follows a normal distribution with mean 0 and variance 1/T where T indicates number of observations in the series. Then 95% confidence interval for ρk is given by

$$\hat{\rho}_k \pm 1.96SE(\hat{\rho}_k) = \hat{\rho}_k \pm 1.96\sqrt{\frac{1}{T}}$$

Decision rule: When ρk falls outside this interval, we reject the null hypothesis that ρk = 0 and conclude that Yt series is non-stationary. On the other hand, if ρk falls within this interval, we accept the null hypothesis that means Yt series is stationary.

4. Box-Pierce Test: This test is applied to examine the validity of the null hypothesis that all ρk's are simultaneously/at same time/concurrently 0. The Box-Pierce Q-statistic may be computed as under:

$$Q^*_{BP} = T\sum_{k=1}^{m} \hat{\rho}_k^2$$

Where, T means number of observations and m means maximum lag length.

Here, Q statistic follows chi-square distribution with m degrees of freedom. So when Computed

$$Q^*_{BP} = \chi^{2*} > \chi^2$$

at the chosen significance level and given degrees of freedom, reject null hypothesis that all ρk are simultaneously 0, and conclude series is non-stationary.

5. Ljung-Box Test: Actually, this test is a modified version of the Box-Pierce test. The modified Ljung-Box Q-statistic is computed as under:

$$Q^*_{LB} = T(T+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{T-k}$$

Here

Q^*_{LB}

Also follows Chi-square distribution with m degrees of freedom and the decision rule is same as Box-Pierce test.

6. Unit Root test: A more formal test of stationarity that has become widely popular is the unit root test. To understand this test you have to consider first an autoregressive (AR) function/model:

$$\mathbf{Y_t = \rho Y_{t-1} + \mu_t}$$

Where μt is the white noise error term such that

$E(\mu_t) = 0$ for all t

$E(\mu_t^2) = \sigma 2$ for all t

$E(\mu_t\ \mu_s) = 0$ but $t \neq s$

We know that when ρ = 1, we face a non-stationary situation and conclude that Yt has a unit root. This also implies that Yt represents a random walk series. Therefore, one way to test if Yt is non-stationary is to regress it on its one period lagged value, i.e., Yt-1, and find out if $\hat{\rho}$ is statistically significantly equal to 1. If it is so, we conclude that Yt is non-stationary.

Alternatively, we may rewrite the above equation as under:

In this situation we cannot apply conventional t-test here for the validity of the null hypothesis (H0): δ = 0. If δ = 0, then ρ = 1, which means that Yt has a unit root and it is non-stationary.

To solve this problem we can apply Dickey-Fuller (DF) or Augmented Dickey-Fuller (ADF) test.

7. Dickey-Fuller test (1979): In practice D-F test is applied in three forms:

$$\Delta Y_t = \delta Y_{t-1} + \mu_t \text{ --- (when the time series has stochastic trend but no drift)}$$

$$\Delta Y_t = \alpha + \delta Y_{t-1} + \mu_t \text{ ----- (when time series has both stochastic trend and drift, this model fits into the financial time series such as interest rates and exchange rates)}$$

$$\Delta Y_t = \alpha + \beta_t + \delta Y_{t-1} + \mu_t \text{ ----- (when the series has everything like drift, deterministic trend and stochastic trend and this model fits into trending time series like asset prices or the levels of macroeconomic aggregates like real GDP etc.)}$$

Where, α is a constant (drift) and t is a time trend. For these entire models the null hypothesis is: H0: δ = 0 (Presence of unit-root or non-stationary)

8. Augmented Dickey-Fuller (ADF) test: The ADF test is same as D-F test, except the D-F regression equations are augmented by including m lags of the dependent variable (ΔYt) to correct serial correlation problem in the disturbance term. Here, the null hypothesis (H0) is same like D-F test. Here also three models are considered as before:

$$\Delta Y_t = \delta Y_{t-1} + \sum_{i=1}^{m} \lambda_i \Delta Y_{t-i} + \mu_t$$

Now what is our decision criteria: If the absolute test statistic (ADF test statistic follows t distribution) is more than the critical value (absolute) then we reject the null hypothesis and accept alternative hypothesis.

H0: Time series has unit root or non-stationary

Ha: Time series has no unit root or stationary

For example: Computed test statistics = 2.0671 and critical value at 5% level is 3.020686

Here, test statistic is less than the critical value (2.0671 < 3.020686), so we cannot reject null hypothesis meaning that time Yt has unit root or non-stationary.

Or

You can check it by using the probability value. If the computed probability value corresponding to the test statistics is higher than the chosen significance level say 5% then reject H0 and vice-versa.

Regression analysis based on time series data implicitly assumes that the underlying time series are stationary. The classical t tests, F tests, etc. are based on this assumption. In practice most economic time series are nonstationary. A stochastic process is said to be weakly stationary if its mean, variance, and autocovariances are constant over time (i.e., they are time invariant). At the informal level, weak stationarity can be tested by the correlogram of a time series, which is a graph of autocorrelation at various lags. For stationary time series, the correlogram tapers off quickly, whereas for nonstationary time series it dies off gradually. For a purely random series, the autocorrelations at all lags 1 and greater are zero. At the formal level, stationarity can be checked by finding out if the time series contains a unit root. The Dickey–Fuller (DF) and augmented Dickey–Fuller (ADF) tests can be used for this purpose. An economic time series can be trend stationary (TS) or difference stationary (DS). A TS time series has a deterministic trend, whereas a DS time series has a variable, or stochastic, trend. The common practice of including the time or trend variable in a regression model to detrend the data is justifiable only for TS time series.

4.2.2 Autoregressive (AR) process

Autoregressive (AR) models represent one of the most fundamental approaches to modelling time series data in economics. The term "autoregressive" literally means "self-regressing" - the variable regresses on its own past values. This concept captures a crucial economic reality: current economic outcomes are often heavily influenced by past economic states.

Consider unemployment rates. If unemployment is high this month, it's likely to remain elevated next month due to factors like job search frictions, employer hesitancy, and economic momentum. This persistence makes unemployment a natural candidate for AR modelling.

The structure of disturbance term in the autoregressive process (AR) is assumed as

$$u_t = \phi 1 u_{t-1} + \phi 2 u_{t-2} + ... + \phi_q u_{t-q} + \varepsilon_t ,$$

i.e., the current disturbance term depends on the q lagged disturbances and $\phi_1, \phi_2, ..., \phi_k$ are the parameters (coefficients) associated with $u_{t-1}, u_{t-2}, ..., u_{t-q}$ respectively. An additional disturbance term is introduced in u_t which is assumed to satisfy the following conditions:

$$E(\varepsilon_t) = 0$$
$$E(\varepsilon_t \varepsilon_{t-s}) = \begin{cases} \sigma_\varepsilon^2 & \text{if } s = 0 \\ 0 & \text{if } s \neq 0. \end{cases}$$

This process is termed as $AR(q)$ process. In practice, the $AR(1)$ process is more popular.

The number of lags used as regressors is called the order of autoregression. So, the preceding model is a first-order autoregression and it is written as AR(1) where 1 represents the order. Since in time series, we have measured values of a variable over time, therefore, we use "t" as a subscript in the variable of time series models. If Y_t and Y_{t-1} are the values of a variable at time 't 'and 't- 1', respectively then we can express the first-order autoregressive model as follows:

$$y_t = \delta + \phi_1 y_{t-1} + \varepsilon_t$$

The model expresses the present value as a linear combination of the mean of the series δ (read as delta), the previous value of the variable Y_{t-1} and the error term εt (read as epsilon). The magnitude of the impact of the previous value on the present value is quantified using a coefficient denoted with φ (read as phai). The "error term " is called white noise and it is normally distributed with mean zero and constant variance (σ^2).

Mean and Variance

We can find mean of an AR(1) model as follows:

$$E[y_t] = \mu = E[\delta] + \phi_1 E[y_{t-1}] + E[\varepsilon_t]$$

Thus, $\text{Mean} = \delta + \phi_1 \mu + 0$ [Since time series is stationary, therefore, $E[y_t] = E[y_{t-1}] = \mu$ and $\varepsilon_t \sim N[0, \sigma^2]$, therefore, $E[\varepsilon_t] = 0$]

$$\text{Mean} = \frac{\delta}{1 - \phi_1}$$

Similarly,

We can find the variance of an AR(1) model as follows:

$$Var[y_t] = Var[\delta] + \phi_1^2 Var[y_{t-1}] + Var[\varepsilon_t]$$

Since time series is stationary, therefore, $Var[y_t] = Var[y_{t-1}]$

Also, δ is a constant, therefore, $Var[\delta] = 0$ and $\varepsilon_t \sim N[0, \sigma^2]$, therefore,

$$Var[\varepsilon_t] = \sigma^2$$

Therefore,

$$Var[y_t] = \phi_1^2 Var[y_t] + \sigma^2$$

$$Var[y_t] = \frac{\sigma^2}{1 - \phi_1^2} \geq 0 \text{ when } \phi_1^2 < 1$$

The autocorrelation function (ACF) for an AR(1) model is as follows:

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi_1^k \text{for } k = 0,1,2, \ldots$$

Since ϕ_1 lies between –1 to 1, therefore, for a positive value of ϕ_1, the ACF (ϕ_1^k)exponentially decreases to 0 as the lag $k$increases. For the negative value of ϕ_1, the ACF also exponentially decays to 0 as the lag increases, but the algebraic signs for the autocorrelations alternate between positive and negative.

Conditions for Stationarity

An autoregressive model can be used on time series data if and only if the time series is stationary. Therefore, some constraints on the values of the parameters are required for stationarity which are as follows:

$$|\phi_1| < 1 \Rightarrow -1 < \phi_1 < 1$$

Since stationarity is necessary for applying autoregressive models, therefore, before applying it to time series data, you will have to check whether the time series is stationary or not. If it is nonstationary, then you will have to apply some transformation methods (such as differencing, log transformation, etc.) to transform the series into a stationary and we use the ARIMA model.

Second-order Autoregressive Models

The autoregressive model in which the value of a variable in the current period is regressed against its two previous values then the autoregressive model is called the second-order autoregression model.

The second-order autoregression model is written as AR(2) where 2 represents the order.

If Y_t, $Y_{t-1}$and Y_{t-2} are the values of a variable at time t,,t− 1 and t −2, respectively then the second-order autoregressive model is expressed as follows:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

Conditions for Stationarity

Since an autoregressive model can be used on time series data if and only if the time series is stationary. Therefore, some constraints on the values of the parameters are required for stationarity which are as follows:

- $|\phi_2| < 1 \Rightarrow -1 < \phi_2 < 1$
- $\phi_1 + \phi_2 < 1$
- $\phi_2 - \phi_1 < 1$

Mathematical Foundation

The general AR(p) model is:

$$y_t = \alpha + \varphi^1 y_{\{t-1\}} + \varphi^2 y_{\{t-2\}} + \ldots + \varphi_p y_{\{t-p\}} + \varepsilon_t$$

This deceptively simple equation captures complex dynamics. Each coefficient φ_i represents the marginal effect of the i-th lag on the current value. The error term ε_t captures unpredictable shocks - news, policy changes, or other random events that affect the economy.

The AR(1) model is:

$$y_t = \alpha + \varphi^1 y_{\{t-1\}} + \varepsilon_t$$

This model embodies the concept of **economic persistence**. The parameter φ_1 measures how much of yesterday's value carries over to today. In economic terms:

- If $\varphi_1 = 0.8$, then 80% of any shock persists to the next period
- If $\varphi_1 = 0.3$, shocks die out more quickly
- If $\varphi_1 = 0.95$, the series exhibits very high persistence

Persistence in economics is crucial for policy analysis. Highly persistent series (φ_1 close to 1) respond slowly to policy interventions, while less persistent series adjust more quickly. This has profound implications for:

Monetary Policy: If inflation is highly persistent, central banks need to act more aggressively to achieve their targets. The famous "Taylor Rule" implicitly accounts for inflation persistence.

Fiscal Policy: If unemployment is highly persistent, temporary fiscal stimulus may have limited long-term effects. Understanding persistence helps design effective intervention strategies.

Business Cycle Analysis: Persistent variables contribute to business cycle propagation. A temporary shock can have lasting effects if it affects persistent variables like employment or investment.

One of the most important tools in AR analysis is the impulse response function. For an AR(1) model, the response to a one-unit shock at time t is:

- Period t: 1 unit
- Period t+1: φ_1 units
- Period t+2: φ_1^2 units
- Period t+k: φ_1^k units

This geometric decay pattern shows how shocks propagate through the economy. The **half-life** of a shock (time for effect to reduce by half) is $\ln(0.5)/\ln(\varphi_1)$, providing a concrete measure of persistence.

4.2.3 Moving average (MA) process

The structure of disturbance term in the moving average (MA) process is

$$u_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots\ldots\ldots + \theta_p \varepsilon_{t-p},$$

i.e., the present disturbance term u_t depends on the p lagged values. The coefficients $\theta_1, \theta_2, \ldots, \theta_p$ are the parameters and are associated with $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-p}$ respectively. This process is termed as $MA(p)$ process.

Moving average (MA) models are the models in which the value of a variable in the current period is regressed against the residuals in the previous period.

A moving average model, states that the current value is linearly dependent on the past error terms.

Conditions for Invertibility

The moving average models are always stationary. However, some restrictions are also imposed on the parameters of the moving average models otherwise the model cannot converge. Therefore, some constraints on the values of the parameters are required for the invertibility of the MA(1) model which is as follows:

$$|\theta_1| < 1 \Rightarrow -1 < \theta_1 < 1$$

Second order Moving Average Models

The moving average model in which the value of a variable in the current period is regressed against its two previous residuals then it is called the second-order moving average. It is represented by MA(2). For example, if today's price of a share depends on whatever has happened in the other factor in the previous day and the day before the previous day then we use a second-order moving average.

4.2.4 Joint Autoregressive Moving Average (ARMA) process

The structure of disturbance term in the Joint Autoregressive Moving Average (*ARMA*) process is

$$u_t = \phi_1 u_{t-1} + \ldots + \phi_q u_{t-q} + \varepsilon_1 + \theta_1 \varepsilon_{t-1} + \ldots + \theta_p \varepsilon_{t-p}.$$

This is termed as $ARMA(q, p)$ process.

The method of correlogram is used to check that the data is following which of the processes.

The autocorrelation function begins at some point determined by both the *AR* and *MA* components but thereafter, declines geometrically at a rate determined by the *AR* component.

In general, the autocorrelation functionis nonzero but is geometrically damped for *AR* process. becomes zero after a finite number of periods for *MA* process.

The *ARMA* process combines both these features.

The results of any lower order of process are not applicable in higher-order schemes. As the order of the process increases, the difficulty in handling them mathematically also increases.

Autoregressive moving average models are simply a combination of an AR model and an MA model. Autoregressive Moving Average (ARMA) models are models in which the value of a variable in the current period is related to its own values in the previous period as well as values of the residual in the previous period.

Since ARMA is a combination of both autoregressive terms(p) and moving average(q) terms, therefore, we represent it as ARMA (p,q). It is also used for stationary time series.

Various Forms of ARMA Models

The ARMA model has various forms for different values of the parameters p and q of the model. We discuss some standard forms as follows:

ARMA(1,1) Models

ARMA(1,1) models are the models in which the value of a variable in the current period is related to its own value in the previous period as well as values of the residual in the previous period. It is a mixture of AR(1) and MA(1).

If Yt and Yt-1are the values of a variable at time t and t –1, respectively and if ε_t and ε_{t-1} are the residuals at time t and t –1, respectively then the ARMA (1,1) model is expressed as follows:

$$y_t = \delta + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

As usual, the coefficients δ and ε t denote the intercept/constant factor and error term at time t, respectively whereas the coefficients φ1 and θ1 represent AR and MA coefficients and represent the magnitude of the impact of past values and past error on the present value, respectively.

The autocorrelation function of an ARMA(1, 1) model exhibits exponential decay and/or sinusoid pattern towards zero. It does not cut off but gradually decreases as lag k increases. Also the autocorrelation function of an ARMA(1,1) model displays the shape of an AR(1) process. The partial autocorrelation function of an ARMA(1, 1) model also gradually dies out (the same property as a moving average model) as k increases. It is relatively difficult to the identification of the order of the ARMA model.

4.2.5 Autoregressive Integrated Moving Average Models

Time series modes such as autoregressive (AR), moving average (MA) and autoregressive moving average (ARMA) models are based on the assumption that the time series is stationary. But in the real world, most of the time series variables are nonstationary. In general, trends, and periodicity exist in many time series data. Hence, the AR, MA, and ARMA models do not apply to nonstationary time series so there is a need to remove these effects before applying such models. Therefore, if the input time series is nonstationary, then first we have to transform the series from a nonstationary into a stationary and after that, we shall apply models such as the AR, MA, and ARMA. For transforming a nonstationary time series to stationary, we may use differencing, once, twice or three times, and so on until the series is at least approximately stationary. As AR and MA processes are described by the order, in a similar way, the differencing process is also described by the order of differencing, as 1, 2, 3.... Therefore, to describe a model for nonstationary time series, the elements make up a triple (p,d,q) instead of two (p, q) that defines the type of model applied where the degree of the differencing is represented by the d parameter. Combining the differencing of a nonstationary time series with the ARMA model provides a powerful family of models that can be applied in a wide range of situations. The model is described as an autoregressive moving average (ARMA) model. In this form, the letter "I" in ARIMA refers to the fact that the time series data has been initially differenced and when the modelling is completed the results then have to be summed or integrated to produce the final estimations and forecasts. Box and Jenkins played a significant role in the development of this extended variant of the model, therefore, ARIMA models are also referred to as Box-Jenkins models.

Autoregressive Integrated Moving Average (ARIMA) model is a combination of differencing with autoregressive and moving average models.

We can express the ARIMA model as follows:

$$y'_t = \delta + \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \ldots + \phi_p y'_{t-q} - \theta_1 \varepsilon_1 - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where $y'_t$is the differenced series which may have been differenced more than once and $p$and q are the orders of autoregressive and moving average parts.

Summarised Overview

Correctly diagnosing and modelling unit roots is fundamental in econometric practice. Ignoring non-stationarity may result in unreliable models and erroneous inferences, whereas proper unit root testing and the use of ARIMA-type models ensure analyses are robust, meaningful, and relevant for both predictive and structural interpretation. These building blocks not only prevent methodological pitfalls like spurious regressions but also open the way to advanced modelling techniques and policy-relevant forecasting in economics.

Assignments

1. What is a unit root and why is it important to test for unit roots in time series analysis?
2. Describe the difference between an AR (Autoregressive) model and an MA (Moving Average) model.
3. Write down the general form of an AR(1) model and specify the condition for stationarity.
4. Suppose you estimate an ARMA(1,1) model and find significant autocorrelation remaining in the residuals. What next diagnostic step would you take?
5. Outline a step-by-step procedure for selecting an appropriate ARIMA model for a monthly sales time series.

Reference

1. Walter Enders (2013) *Applied Econometric Time Series*, 3rd Edition, Wiley, ISBN-10: 8126543914, ISBN-13: 978-8126543915.
2. Jeffrey M. Wooldridge (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition, MIT Press, ISBN-10: 9384106259, ISBN-13: 978-9384106256
3. James D Hamilton *Time Series Analysis* (1994) Princeton University Press

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.
2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

UNIT 3

Panel Data Models: Fixed and Random Effects

Learning Outcomes

After completing this unit, the learner will be able to:

- understand the Structure and Advantages of Panel Data
- estimate and Interpret Fixed Effects Models (Including LSDV)
- estimate and Interpret Random Effects Models
- distinguish Between Fixed Effects and Random Effects Models

Background

Modern economic and social research frequently relies on datasets that track multiple subjects over time - such as firms, countries, or individuals observed annually or monthly. These two-dimensional datasets, known as panel data, offer unique advantages over simple cross-sectional or time series data by allowing researchers to control for unobserved, time-invariant differences between entities, uncover dynamic relationships, and generate more credible causal estimates.

In this unit, the estimation and interpretation of panel data regression models take center stage. Two principal approaches - the Fixed Effects (FE) model, often implemented through the Least Squares Dummy Variable (LSDV) method, and the Random Effects (RE) model - allow analysts to address individual-specific heterogeneity and decide the most appropriate structure for their research questions.

Keywords

Fixed Effects Model (FE), Least Squares Dummy Variable (LSDV) Model, FE approach, Random Effects Model (RE)

Discussion

4.3.1 Panel Data

Panel data, also known as longitudinal data, combines both cross-sectional and time-series dimensions. This rich data structure allows economists to control for unobserved heterogeneity and provides more robust econometric analysis. Understanding panel data regression models is crucial for modern empirical economics, as they help address many limitations inherent in pure cross-sectional or time-series analyses.

Panel data consists of observations on multiple entities (individuals, firms, countries) over multiple time periods. We denote this as:

$$y_{it} = f(x_{it}, \alpha_i, \lambda_t, \varepsilon_{it})$$

Where:

$i = 1, 2, ..., N$ (cross-sectional units)

$t = 1, 2, ..., T$ (time periods)

y_{it} = dependent variable for unit i at time t

x_{it} = vector of explanatory variables

α_i = individual-specific effects

λ_t = time-specific effects

ε_{it} = error term

Panel Data: An Illustrative Example

Following analysis was carried out based on the data taken from a famous study of investment theory proposed by Grunfeld (1958).

Grunfeld was interested in finding out how real gross investment (Y) depends on the real value of the firm (X_2) and real capital stock (X_3). It includes data on four companies, General electric (GE), General Motor (GM), U.S. Steel (US), and Westinghouse. Data for each company on the preceding three variables are available for the period 1935-54. Thus, there are four cross-sectional units and 20 time periods. In all, therefore, we have 80 observations. A prior, Y is expected to be positively related to X_2 and X_3.

Pooling, or combining, all the 80 observations, the Grunfeld investment function can be written as:

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it} \quad \text{.....(1)}$$

$i = 1, 2, 3, 4$ and $t = 1, 2, . . ., 20$

where i stands for the i^{th} cross-sectional unit and t for the t^{th} time period and it is assumed that

the X's are nonstochastic and that the error term follows the classical assumptions, namely, E(unit) ~ N(0, σ^2) .

4.3.2 Estimation of panel data regression models

Fixed Effects Approach:

Estimation depends on the assumptions we make about the intercept, the slope coefficients, and the error term. There are several possibilities:

- Assume that the intercept and slope coefficients are constant across time and space and the error term captures differences over time and individuals.
- The slope coefficients are constant but the intercept varies over individuals.
- The slope coefficients are constant but the intercept varies over individuals and time.
- All coefficients (the intercept as well as slope coefficients) vary over individuals.
- The intercept as well as slope coefficients vary over individuals and time.

i. All coefficients constant across time and individuals

The simplest, and possibly naive approach is to disregard the space and time dimensions of the pooled data and just estimate the usual OLS regression. That is, stack the 20 observations for each company one on top of the other, thus giving in all 80 observations for each of the variables in the model.

The OLS results are as follows:

$$\hat{Y} = -63.3041 + 0.1101X_2 + 0.3034X_3$$

se = (29.6124) (0.0137) (0.0493)

t = (2.1376) (8.0188) (6.1545)

$R^2 = 0.7565$ Durbin-Watson = 0.2187, n = 80 df = 77 ..... (2)

Here all the coefficients are individually statistically significant and the R^2 value is reasonably high. But the only problem seems to be the estimated Durbin-Watson statistic which is quite low, suggesting that perhaps there is autocorrelation in the data. The estimated model assumes that the intercept value of GE, GM, US, and Westinghouse are the same. It also assumes that the slope coefficients of two X variables are all identical for all the four firms. Obviously, these are very restricted assumptions. Therefore, despite its simplicity the pooled regression may distort the true picture of the relationship between Y and X's across the four companies.

ii. The slope coefficients are constant but the intercept varies over individuals

Due to the Fixed Effects or Least-Squares Dummy Variables (LSDV) Regression Model

One way to take into account the individuality of each company or each cross-sectional unit is

to let the intercept vary for each company but still assume that the slope coefficients are constant across firms. We write the model as:

$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it}$ ..... (3)

The difference in the intercept may be due managerial style or managerial philosophy.

The model (3) is known as the fixed effects (regression) model (FEM). The term "fixed effects" is due to the fact that, although the intercept may differ across individuals, each individual's intercept does not vary over time; that is, it is time invariant.

This can be done by the dummy variable technique. Therefore, we write the model as

$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it}$ ..... (4)

where $D_{2i} = 1$ if the observation belongs to GM, 0 otherwise;

$D_{3i} = 1$ if the observation belongs to US, 0 otherwise; and $D_{4i} = 1$

if the observation belongs to WEST, 0 otherwise. Here α_1 represents the intercept of GE and $\alpha_2, \alpha_3, \text{and} \alpha_4$, the differential intercept coefficients, tell by how much the intercepts of GM, US, and WEST differ from the intercept of GE. Since we are using dummies to estimate the fixed effects, the model is also known as the least-squares dummy variable (LSDV) model. The results are as follows;

$\hat{Y}_{it} = -245.7924 + 161.5722 D_{2i} + 339.6328 D_{3i} + 186.5666 D_{4i} + 0.1079 X_{2i} + 0.3461 X_{3i}$

se = (35.8112) (46.4563) (23.9863) (31.5068) (0.0175) (0.0266)

t = (6.8635) (3.4779) (14.1594) (5.9214) (6.1653) (12.9821)

$R^2 = 0.9345$ d=1.1076 df =74 ..... (5)

Here all the estimated coefficients are individually highly significant and the intercept values of the four companies are statistically different. The differences in the intercepts may be due to unique features of each company, such as differences in management style or managerial talent. Judged by the statistical significance of the estimated coefficients, and the fact that the R^2value has increased substantially we can conclude that (5) is better than (2). The Durbin-Watson *d* value is much higher; suggesting that model (2) was miss-specified.

We can also provide a formal test of the two models. In relation to (5), model (2) is a restricted model in that it imposes a common intercept on all the companies. Therefore, we can use the restricted F test. Using the formula, we get

$$F = \frac{(R^2_{UR} - R^2_R)/3}{(1 - R^2_{UR})/74} = 66.9980$$

where the restricted value is from (2) and the unrestricted is from (5).

Clearly, the F value of 66.998 is highly significant and, therefore, the restricted regression (2) seems to be invalid.

4.3.3 The Time Effect

Just as we used the dummy variables to account for individual effect, we can allow for time effect in the sense that the Grunfeld investment function shifts over time. For such a situation we introduce time dummies, one for each year. From the regression results, we infer that none of the individual time dummies were individual statistically significant. We have already seen that the individual company effects were statistically significant, but the individual year effects were not.

i. Slope coefficients constant but the intercept varies over individual as well as time

To consider this possibility, we can combine (5) and the time effect model, as follows:

$$Y_{it}=\alpha_1+\alpha_2 D_{GMi}+\alpha_3 D_{USi}+\alpha_4 D_{WESTi}+\lambda_0+\lambda_1 DUM35+\ldots+\lambda_{19} DUM53+\ldots+\beta_2 X_{2i}+\beta_3 X_{3i}+u_{it} \quad \ldots\ldots\ldots \quad (7)$$

when we run this regression, we find the company dummies as well as the coefficients of the X are individually statistically significant, but none of the time dummies are. Essentially, we are back to (5).

ii. All coefficients vary across individuals

Here we assume that the intercepts and the slope coefficients are different for all individual, or cross-section, units. This is to say that the investment functions of GE, GM, US, and WEST are all different. We can easily extend our LSDV model to take care of this situation. Here what we do is multiply each of the company dummies by each of the X variables. That is, we estimate the following model:

$$Y_{it}=\alpha_1+\alpha_2 D_{2i}+\alpha_3 D_{3i}+\alpha_4 D_{4i}+\beta_2 X_{2it}+\beta_3 X_{3it}+\gamma_1(D_{2i}X_{2it})+\gamma_2(D_{2i}X_{3it})+\gamma_3(D_{3i}X_{2it})+\text{----}+uit \quad \ldots\ldots. \quad (8)$$

The γ's are the differential slope coefficients, just as $\alpha_2, \alpha_3,$ and α_4 are the differential intercepts. If one or more of the γ coefficients are statistically significant, it will tell us that one or more slope coefficients are different from the base group. If all the differential intercept and all the differential slope coefficients are statistically significant, we can conclude that the investment functions are different for the four companies.

Estimation of panel date regression models: The Random Effects Approach

Although fixed effects or LSDV model can be expensive in terms of degrees of freedom if we have several cross-sectional units. If the dummy variables do in fact represent a lack of knowledge about the (true) model, why not express this ignorance through the disturbance term υ_{it} ? This is precisely the approach suggested by the proponents of the so called error components model (ECM) or random effects model (REM).

The basic idea is to start with (3):

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \upsilon_{it} \quad \ldots\ldots\ldots\ldots \quad (9)$$

Instead of treating β_{1i} as fixed, we assume that it is a random variable with a mean value of β_1. And the intercept value for an individual company can be expressed as

$\beta_{1i} = \beta_1 + \varepsilon_i \qquad i = 1, 2, \ldots, N \qquad \ldots\ldots\ldots\ldots (10)$

where ε_i is a random error term with a mean value of zero and variance σ^2.

The four firms included in the sample are a drawing from a much larger universe of such companies and they have a common mean value for the intercept and the individual differences in the intercept values of each company are reflected in the error term.

Substituting (10) into (9), we get:

$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + \upsilon_{it}$

$Y_{it} = Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \omega_{it} \qquad \ldots\ldots\ldots\ldots (11)$

where

$\omega_{it} = \varepsilon_i + \upsilon_{it} \qquad \ldots\ldots\ldots\ldots (12)$

The composite error term consists of two components, the cross-section, or individual-specific, error component and the combined time series and cross-section error component. The usual assumptions made by ECM are that

$\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$

$\upsilon_{it} \sim N(0, \sigma^2_\upsilon)$

$E(\varepsilon_i \upsilon_{it}) = 0 \qquad E(\varepsilon_i \varepsilon_j) = 0 \qquad (i \neq j)$

$E(\upsilon_{it} \upsilon_{is}) = E(\upsilon_{it} \upsilon_{jt}) = E(\upsilon_{it} \upsilon_{js}) = 0 \quad (i \neq j; t \neq s) \qquad \ldots\ldots\ldots\ldots (13)$

That is, the individual error components are not correlated with each other and are not auto correlated across both cross-section and time series units.

However, it can be shown that ω_{it} and ω_{is} are correlated; that is, the error terms of a given cross-sectional unit at two different points in time are correlated.

If we do not take this correlation structure into account, and estimate (11) by OLS, the resulting estimators will be inefficient. The most appropriate method here is the method of Generalized least squares (GLS).

4.3.4 Fixed Effects versus Random Effects Model

Aspect	**Random Effects**	**Fixed Effects**
Treatment of α_i	Random variable	Fixed parameter
Key Assumption	Cov(α_i, X_{it}) = 0	No assumption needed
Estimation	GLS	Within transformation

Aspect	Random Effects	Fixed Effects
Time-invariant variables	Can be estimated	Cannot be estimated
Degrees of freedom	More efficient	Loses N degrees of freedom

Panel data models examine group (individual-specific) effects, time effects, or both. These effects are either fixed effect or random effect. A fixed effect model examines if intercepts vary across groups or time periods, whereas a random effect model explores differences in error variances. A one-way model includes only one set of dummy variables (*e.g.,* firm), while a two-way model considers two sets of dummy variables (*e.g.,* firm and year). If it is assumed that the error component and the X's are uncorrelated, ECM may be appropriate, whereas if they are correlated, FEM may be appropriate. Keeping this fundamental difference in the two approaches in mind, the choice between FEM and ECM may be done by:

- If T (the number of time series data) is large and N (the number of cross-sectional units) is small, there is likely to be little difference in the values of the parameters estimated by FEM and ECM. Hence the choice here is based on computational convenience. On this score, FEM may be preferable.
- When N is large and T is small, the estimates obtained by the two methods can differ significantly. in ECM
- $\beta_{1i} = \beta_1 + \varepsilon_i$, but in FEM we treat β_{1i} as fixed and non-random.
- If the individual error component and one or more repressors
- are correlated, then the ECM estimators are biased, whereas those obtained from FEM are unbiased.
- If N is large and T is small, and if the assumptions underlying ECM hold, ECM estimators are more efficient than FEM estimators.

Hausman test (1978) is used to choose between FEM and ECM. The null hypothesis underlying the Hausman test is that the FEM and ECM estimators do not differ substantially. The test statistic developed by Hausman has an asymptotic chi-square distribution. If the null hypothesis is rejected, the conclusion is that ECM is not appropriate and that we may be better off using FEM, in which case statistical inferences will be conditional on the error component in the sample.

Panel data, by blending the inter-individual differences and intra-individual dynamics have advantages over cross-sectional or time-series data. It has greater capacity for capturing the complexity of human behavior than a single cross-section or time series data. More accurate inference of model parameters can be obtained through panel data. Panel data usually contain more degrees of freedom and more sample variability than cross-sectional data or time series. It controls the impact of omitted variables, i.e., reduces omitted variable bias. Panel data helps in uncovering dynamic relationships.

The choice between fixed effects and random effects models represents one of the most fundamental decisions in panel data econometrics. This decision is not merely technical—it embodies different philosophical approaches to modeling unobserved heterogeneity and carries profound implications for the interpretation and validity of empirical results. While both approaches seek to address the challenge of unobserved individual-specific effects, they do so through fundamentally different assumptions and methodologies.

This comprehensive explanation provides an in-depth comparison of these two approaches, examining their theoretical foundations, estimation procedures, statistical properties, and practical applications. Understanding when and why to use each approach is crucial for conducting credible empirical research in economics, where the treatment of unobserved heterogeneity can make or break the validity of policy conclusions.

The Hausman test provides statistical guidance for this choice, but economic theory and institutional knowledge should ultimately guide the decision. In practice, many researchers estimate both models as robustness checks, with the understanding that fixed effects provides a "worst-case" scenario for identification while random effects provides a "best-case" scenario for efficiency.

The evolution of panel data methods continues to blur the lines between these approaches. Correlated random effects, Hausman-Taylor estimators, and hybrid models represent attempts to capture the benefits of both approaches while minimizing their respective drawbacks. As panel datasets become larger and more complex, these intermediate approaches are likely to become increasingly important.

Understanding the trade-offs between fixed and random effects is crucial for any empirical economist. This choice affects not only the statistical properties of your estimates but also the substantive conclusions you can draw from your research. The careful consideration of these trade-offs, guided by both statistical tests and economic reasoning, represents the art of applied econometrics at its finest.

Remember that no single approach is universally superior. The best choice depends on your research question, data characteristics, and the plausibility of underlying assumptions. Master both approaches, understand their strengths and limitations, and choose thoughtfully based on your specific empirical context.

Summarised Overview

This unit demonstrated how panel data regression models - specifically the fixed effects (FE) and random effects (RE) approaches - empower researchers to analyze data that spans both entities and time. By synthesizing cross-sectional and time-series perspectives, panel data models address unobserved heterogeneity, reduce omitted variable bias, and enable richer, more accurate investigations of dynamic causal relationships.

Assignments

1. Explain the rationale behind using the Fixed Effects (FE) Model for panel data. What type of unobserved heterogeneity does it address?
2. Describe the Least Squares Dummy Variable (LSDV) approach to estimating fixed effects. In what situations is this approach equivalent to other FE methods?
3. What are the core assumptions of the Random Effects (RE) model? How is it different from the Fixed Effects model?
4. Under what circumstances would the Random Effects model be more efficient than Fixed Effects?

Reference

1. Allison, P. D. (2009). *Fixed Effects Regression Models*. Sage Publications.
2. Baltagi, B. H. (2021). *Econometric Analysis of Panel Data* (6th ed.). Springer.
3. Baltagi, B. H., & Li, Q. (1991). *A Monte Carlo Study of the Hausman Test for Correlated Random Effects*. Economics Letters, 37(2), 115-119.
4. Hausman, J. A. (1978). *Specification Tests in Econometrics*. Econometrica, 46(6), 1251-1271.
5. Hsiao, C. (2022). *Analysis of Panel Data* (4th ed.). Cambridge University Press.
6. Kiviet, J. F. (1995). On bias, *Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models*. Journal of Econometrics, 68(1), 53-78.
7. Mundlak, Y. (1978). *On the Pooling of Time Series and Cross Section Data*. Econometrica, 46 (1), 69-85.
8. Plümper, T., & Troeger, V. E. (2007). *Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects*. Political Analysis, 15(2), 124-139.

9. Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics* (4th ed.). Pearson.

10. Wooldridge, J. M. (2019). *Introductory Econometrics : A Modern Approach* (7th ed.). Cengage Learnin

Suggested Reading

1. Damodar N Gujarati and Dawn C Porter (2011*):* *Econometrics by Example*, Palgrave Macmillan; New edition.

2. G.S.Maddala (1992): *Introduction to Econometrics*, Second Edition, Macmillan Publishing Company New York

Space for Learner Engagement for Objective Questions

Learners are encouraged to develop objective questions based on the content in the paragraph as a sign of their comprehension of the content. The Learners may reflect on the recap bullets and relate their understanding with the narrative in order to frame objective questions from the given text. The University expects that 1 - 2 questions are developed for each paragraph. The space given below can be used for listing the questions.

SREENARAYANAGURU OPEN UNIVERSITY

QP CODE: …… Reg. No : .....

Name : …….....

Model Question Paper- set-I

FOURTH SEMESTER MA ECONOMICS EXAMINATION
DISCIPLINE CORE – 11

M23EC11DC - ECONOMETRICS
(CBCS - UG)
2024-25 - Admission Onwards

Time: 2 Hours Max Marks: 45

Section A - Objective Type Questions

Answer any 10 questions. Each question carries 1 mark (10 X 1=10 Marks)

1. With which property is the problem of non-stationarity in a time series primarily associated?
2. Which aspect of OLS estimation is primarily affected by multicollinearity?
3. Which set of assumptions includes the condition that the error term has zero mean?
4. Which OLS assumption is violated in the presence of heteroscedasticity?
5. What values can a dummy variable take?
6. What does ARIMA stand for?
7. What does econometrics primarily deal with?
8. In which type of data does the problem of autocorrelation mostly arise?
9. What do ANOVA and ANCOVA models involve?
10. What does regression through the origin imply?
11. What does the Durbin–Watson test detect?
12. What does stationarity in a time series mean?
13. What type of data is used in econometric analysis?
14. When does the dummy variable trap occur in a regression with dummy variables?
15. What do model specification errors lead to?

Section B- Very Short Questions

Answer any 5 questions. Each question carries 2 marks (5X2=10 Marks)

16. Explain the meaning of heteroscedasticity in a regression model.
17. Define econometrics and explain its scope briefly.
18. What are the consequences of multicollinearity in an OLS model?
19. What is meant by population regression function?
20. What is a dummy variable? Give one example of its use.
21. What are the main limitations of econometric methods?
22. Mention any two tests for detecting autocorrelation.
23. State two assumptions of the Classical Linear Regression Model (CLRM).
24. What do you mean by stationarity in a time series?
25. Define model specification error and give one example.

Section C- Short Answer

Answer any 5 questions. Each question carries 4 marks. (5X4=20 Marks)

26. Distinguish between stationary and non-stationary time series.
27. Explain the methodology of econometrics with suitable steps.
28. Discuss the concept of unit root tests and their importance in time series analysis.
29. Explain the functional forms of regression models with examples.
30. What are the consequences of heteroscedasticity in regression analysis?
31. Explain the use of dummy variables in regression and illustrate with an example.
32. Describe the tests for detecting multicollinearity.
33. Differentiate between ANOVA and ANCOVA models.

Section D- Long Answer/Essay Question

Answer any 3 questions. Each question carries 10 marks. (3X10=30 Marks)

34. Describe in detail the Ordinary Least Squares (OLS) method of estimation and explain the Gauss-Markov theorem.
35. Explain the structure and estimation of a three-variable econometric model and discuss the interpretation of coefficients.
36. Critically examine the violations of CLRM assumptions—heteroscedasticity, multicollinearity, and autocorrelation—with consequences and remedial measures.
37. Discuss the Fixed Effects and Random Effects Models in panel data analysis and bring out their differences.
38. Explain the working of Qualitative Response Models (LPM, Logit, Probit, Tobit) and their applications.
39. Explain the ARIMA model and outline the procedure for time series forecasting.

SREENARAYANAGURU OPEN UNIVERSITY

QP CODE: …… Reg. No : .....

Name : …….....

Model Question Paper- set-II

FOURTH SEMESTER MA ECONOMICS EXAMINATION
DISCIPLINE CORE – 11

M23EC11DC - ECONOMETRICS

(CBCS - UG)
2024-25 - Admission Onwards

Time: 2 Hours Max Marks: 45

Section A - Objective Type Questions

Answer any 10 questions. Each question carries 1 mark (10 X 1=10 Marks)

1. What does the population regression function show?
2. Which of the following describes a property of a stationary series?
3. Which aspect of estimation is mainly affected by the presence of autocorrelation?
4. How is econometrics best defined?
5. Why is the adjusted R^2 preferred over R^2?
6. What type of process does the random walk model represent?
7. What does the term AR process refer to?
8. What does a dummy variable represent in a regression model?
9. What does the log–log model imply about the interpretation of coefficients?
10. What is the purpose of the Partial Autocorrelation Function (PACF) in time series analysis?
11. Which features does the ANCOVA model combine?
12. When does underfitting of a model occur?
13. What are the main characteristics of a white noise process?
14. Which models are suitable when the dependent variable is binary (0,1)?
15. What is the consequence of a specification error in a regression model?

Section B- Very Short Questions

Answer any 5 questions. Each question carries 2 marks (5X2=10 Marks)

16. What is meant by a unit root?
17. What does the coefficient $\beta 1$ measure in a multiple regression model?
18. What is meant by a deterministic trend?
19. What does the Logit model estimate?
20. What does "MA" stand for in time series models?
21. What is the main objective of qualitative response models?
22. What does ARIMA represent?
23. Mention two remedial measures for heteroscedasticity.
24. List two advantages of using functional form transformations.
25. Define multiple correlation coefficient (R^2).

Section C- Short Answer

Answer any 5 questions. Each question carries 4 marks. (5X4=20 Marks)

26. Explain the scope and significance of econometrics in policy formulation.
27. Discuss the problems arising from heteroscedasticity in regression.
28. Explain the importance of dummy variables in multiple regression models.
29. Explain the concept of stationarity and why it matters in time series analysis.
30. Distinguish between qualitative explanatory variables and qualitative response models.
31. Explain the role of ARIMA models in economic forecasting.
32. Discuss the remedial measures for autocorrelation.
33. Explain the meaning of a population regression function.

Section D- Long Answer/Essay Question

Answer any 3 questions. Each question carries 10 marks. (3X10=30 Marks)

34. Examine the violations of CLRM assumptions and their implications for OLS estimation.
35. Explain in detail the estimation of a simple and multiple regression model using OLS.
36. Describe the dummy variable approach in regression and discuss its interpretation.
37. Discuss the methodology and applications of econometric models in empirical research.
38. Discuss the Fixed and Random Effects Models in panel data analysis.
39. Explain model specification and the consequences of specification errors.

സർവ്വകലാശാലാഗീതം

വിദ്യയാൽ സ്വതന്ത്രരാകണം
വിശ്വപൗരരായി മാറണം
ഗ്രഹപ്രസാദമായ് വിളങ്ങണം
ഗുരുപ്രകാശമേ നയിക്കണേ

കൂരിരുട്ടിൽ നിന്നു ഞങ്ങളെ
സൂര്യവീഥിയിൽ തെളിക്കണം
സ്നേഹദീപ്തിയായ് വിളങ്ങണം
നീതിവൈജയന്തി പാറണം

ശാസ്ത്രവ്യാപ്തിയെന്നുമേകണം
ജാതിഭേദമാകെ മാറണം
ബോധരശ്മിയിൽ തിളങ്ങുവാൻ
ജ്ഞാനകേന്ദ്രമേ ജ്വലിക്കണേ

കുരീപ്പുഴ ശ്രീകുമാർ

SREENARAYANAGURU OPEN UNIVERSITY

Regional Centres

Kozhikode
Govt. Arts and Science College
Meenchantha, Kozhikode,
Kerala, Pin: 673002
Ph: 04952920228
email: rckdirector@sgou.ac.in

Thalassery
Govt. Brennen College
Dharmadam, Thalassery,
Kannur, Pin: 670106
Ph: 04902990494
email: rctdirector@sgou.ac.in

Tripunithura
Govt. College
Tripunithura, Ernakulam,
Kerala, Pin: 682301
Ph: 04842927436
email: rcedirector@sgou.ac.in

Pattambi
Sree Neelakanta Govt. Sanskrit College
Pattambi, Palakkad,
Kerala, Pin: 679303
Ph: 04662912009
email: rcpdirector@sgou.ac.in

DON'T LET IT BE TOO LATE

SAY NO TO DRUGS

LOVE YOURSELF AND ALWAYS BE HEALTHY

SREENARAYANAGURU OPEN UNIVERSITY

SREENARAYANAGURU OPEN UNIVERSITY

The State University for Education, Training and Research in Blended Format, Kerala

ECONOMETRICS

COURSE CODE: M23EC11DC

Sreenarayanaguru Open University
Kollam, Kerala Pin- 691601, email: info@sgou.ac.in, www.sgou.ac.in Ph: +91 474 2966841

ISBN 978-81-990500-9-9